MATCHING NUMERICAL SUPERVISION TO THE OUTPUT INTERFACE IN SYMBOLIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural symbolic regressors pass coefficient placeholders or predicted constants to numerical solvers. We show that this output interface can reverse conclusions about numerical supervision. In scale-augmented Flash-ANSR continuation, role targets favor placeholder requests, whereas literal targets favor constant requests. Independent confirmation on 256 newly generated families and three training seeds yields interactions of -6.23 and -7.44 percentage points at $R^2 \geq .99$ and $R^2 \geq .999$, with both multiplicity-adjusted intervals excluding zero. Exploratory paired comparisons show that the interaction varies with continuation exposure. Balanced 50/50 continuation provides a shared-checkpoint training response: it improves the weaker pairings by 3.17 and 3.06 points, reaching observed coverage within 0.70 points of the respective specialists. Controlled interventions distinguish boundary supervision, numerical context, and optimizer initialization. Fixed-history probes locate immediate format steering at coefficient entry; fixed-pool replays show that predicted coefficients carry useful initialization information across candidate expressions. Together, these findings identify the generation–refinement interface as a design axis for numerical supervision and support balanced exposure when both output capabilities are required.

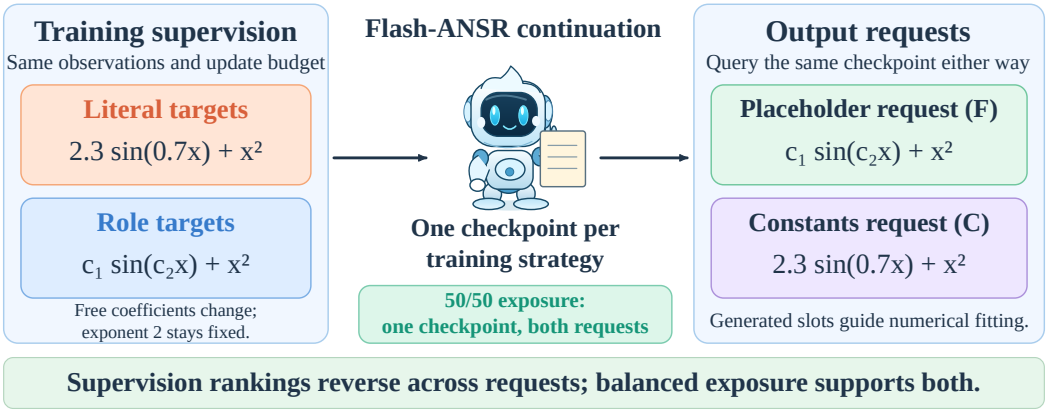


Figure 1: **Numerical supervision and output requests.** Each training strategy produces its own checkpoint; balanced training mixes literal and role targets in equal proportions. Each checkpoint receives both requests. The formulas illustrate the representations: free coefficients change format while exponent 2 stays fixed. Requests condition generation, and emitted slots determine numerical fitting.

1 INTRODUCTION

Neural symbolic regression couples mathematical sequence generation with continuous parameter estimation through two established paradigms: predicting expression skeletons whose free coefficients are fitted entirely by numerical optimizers (Biggio et al., 2021), or generating explicit con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stants that directly initialize local refinement (Kamienny et al., 2022). Because expressions combine free parameters with fixed structural values such as exponents, the training supervision determines which numerical tokens the decoder learns to emit, while the inference interface determines how predictions enter numerical search (Figure 1). Evaluating numerical supervision under a single output interface can select a training objective whose ranking reverses when the request changes.

We examine this interaction in scale-augmented Flash-ANSR continuation (Saegert & Köthe, 2026). The supervision ranking reverses: roles yield higher coverage than literals under placeholder requests, and literals yield higher coverage under constant requests. On 256 newly generated confirmation families and three training seeds, the interaction is -6.23 percentage points at $R^2 \geq 0.99$ (adjusted interval $[-9.84, -2.79]$). At $R^2 \geq 0.999$, it reaches -7.44 points ($[-11.05, -3.86]$). Exploratory paired comparisons at 120M show a negative shift in the interaction under scale training relative to identity continuation (Figure 5b).

Balanced 50/50 continuation trains one checkpoint to support both requests. Alternating target formats improves the weaker pairings by 3.17 and 3.06 percentage points, with positive adjusted intervals and observed coverage within 0.70 points of the respective specialists. Placeholder carryover under constant requests falls from 11.55% to 0.02%. Controlled interventions clarify the roles of numerical information: loss and context interventions distinguish target supervision from teacher history, prefix probes locate immediate steering at coefficient entry, and fixed-parent replays show that predicted coefficients provide useful initial values across candidate expressions. These findings connect the confirmed trade-off to a shared training strategy and specific operations in generation and refinement.

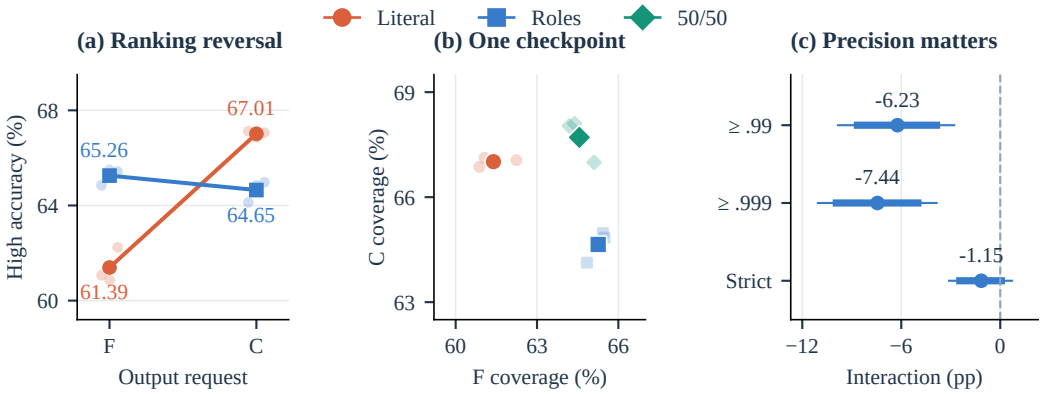


Figure 2: **Ranking reversal, balanced exposure, and independent confirmation.** (a) Cohort B, 256 families and three training seeds; faint points show seeds. (b) Each checkpoint's F and C coverage forms one point; small markers show training seeds. Balanced coverage differs from the respective specialists by -0.69 and $+0.69$ points, with both paired intervals including zero. (c) Interaction $(\text{Roles} - \text{Literal})_C - (\text{Roles} - \text{Literal})_F$: thick intervals are nominal 95%; thin intervals adjust nine comparisons. Figure 9 summarizes capacity and additional-seed comparisons.

2 Related Work

Learning and fitting expressions. NeSymReS predicts skeletons for coefficient fitting (Biggio et al., 2021); E2E and SymFormer predict constants for local refinement (Kamienny et al., 2022; Vastl et al., 2024). Flash-ANSR supplies native parameter roles and a shared generation/refinement stack (Saegert & Köthe, 2026). MMSR jointly trains symbolic and numerical representations (Li et al., 2025); TPSR, SNIP, GenSR, and LEE modify planning or learned search (Shojaee et al., 2023; Meidani et al., 2024; Li et al., 2026; Chu et al., 2026). Our contribution is the crossed target–output evaluation and its balanced training response. Fixed-parent replays separately test initialization.

Interfaces, coordinates, and evaluation. Shared representations and instruction mixtures motivate multi-format exposure (Raffel et al., 2020; Chung et al., 2024); scheduled sampling studies changes in teacher-provided versus generated histories (Bengio et al., 2015). Here, format choices alter the numerical objects received by a solver. Numerical-domain fragility (Voigt et al., 2025) and SRBench/SRSD evaluation (La Cava et al., 2021; Matsubara et al., 2024) motivate separate coverage, risk, and physical-recovery endpoints. ESR and MDLformer connect expression complexity to model selection (Bartlett et al., 2022; Yu et al., 2025). Our S1 selector prices description length in the coordinates used for refinement; algebraic equality is evaluated separately (Kronberger et al., 2026). Appendix AL expands the relation to hybrid search, evolutionary methods, and units-aware SR; Appendix Z covers recent hybrid systems.

3 SUPERVISION, OUTPUT INTERFACES, AND COORDINATES

3.1 TARGET FORMATS AND BALANCED EXPOSURE

Literal targets serialize numbers in the native IEEE-754 format. *Role targets* use `mask_fittable`: designated free coefficients become supervised placeholders, while fixed typed values retain their numerical spelling. In $a\sin(bx)+x^2$, roles replace a and b with placeholders and retain the exponent 2. All retained fixed numeric sites in this corpus are exponents, present in 2,555 of 4,096 training families. Both targets preserve the native floating-target mask while changing the prompt, sequence length, and number of supervised tokens. A global-token averaging control retains the negative interaction in three seeds on 32 families (Appendix Q).

Balanced-format control. Three additional 20M checkpoints use the same initialization, training families, paired scale views, 2,000 updates, and per-view mean loss. Half of the format exposures use the native literal/constants instruction and target; half use role/placeholder instructions and targets. Each resulting checkpoint is then queried under both output directives. This intervention changes format exposure and supervision together (Appendix A).

3.2 FROM NUMERICAL EMISSION TO A RETURNED FORMULA

The native pipeline fits a proposed parameter family $h(x;\theta)$, applies discrete constant refinement, and selects a completed formula. Discrete refinement considers short constant spellings and term deletion. Selection uses only visible observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$:

$$S(f;\mathcal{D}) = \log_{10} \max\{\mathrm{FVU}_{\mathcal{D}}(f), 2^{-52}\} + \frac{2}{n \log_2 10} L(f), \tag{1}$$

where L is length in bits and $\mathrm{FVU}_{\mathcal{D}}$ divides observed mean squared error by observed target variance. Clean labels are reserved for evaluation.

Training targets and output emission are crossed separately (Table 1). Both prompt directives permit placeholders and explicit numbers; each requests an output representation; the candidate's actual refinement slots determine whether a complete predicted p_0 is available. The native policy configures eight Gaussian random starts and an unperturbed predicted start when available; the visible fit criterion determines whether random restarts follow. We compare configurations under a shared refinement policy, with actual work determined by their candidates. Figure 3 traces the emitted coefficient classes and measures the request effect at fixed histories.

Matched initialization. We replay fixed 64-proposal parent pools with eight starts per parameterized candidate. R uses Gaussian starts; P replaces one with the candidate's predicted vector p_0. E uses eight empirical vectors from other expressions, matching the number of free coefficients; PE replaces one empirical vector with p_0 when available. The shared empirical bank combines predictions from two 64-proposal sources, one per training format, primarily from the same query (Appendix AE.2). Parameter ties and typed values remain fixed. Each fit is capped at 200 expression forwards and five seconds; the best finite pre-cap state remains available. The same refinement and visible score select a formula. Appendix AE retains coarse, shuffled, and historical native controls and their actual work.

Table 1: **Crossed configurations and the path from a request to a returned formula.** $B = \langle \text{bos} \rangle$, $E = \langle \text{expression} \rangle$, $M = \langle \text{mask_fittable} \rangle$. F requests placeholders and C constants; both allow both coefficient representations. Actual emitted slots determine the numerical refinement path. Random starts follow a predicted fit when visible FVU exceeds 2^{-23}. Native refinement also fits variants in which fixed typed constants become free parameters. All candidates share discrete refinement and visible selection. Loss is computed over supervised body tokens, whose lengths differ by format.

Setting	Training prefix	Supervised expression body	Inference prefix
Literal / F	$B\,E$	Free and typed numbers spelled numerically	$B\,M\,E$
Literal / C	$B\,E$	Free and typed numbers spelled numerically	$B\,E$
Roles / F	$B\,M\,E$	Free sites are supervised placeholders; typed numbers remain spelled	$B\,M\,E$
Roles / C	$B\,M\,E$	Free sites are supervised placeholders; typed numbers remain spelled	$B\,E$

3.3 Paired scale-augmented continuation

For each base relationship f, we generate one noisy observation set and four views. One is identity; the other three independently sample each input and output scale with $\log_{10}$ factors uniform on $[-6, 6]$. For diagonal input scale S and output factor $t > 0$, the view contains (Sx, ty) and target

$$f_{S,t}(x') = tf(S^{-1}x'). \tag{2}$$

All views share the observations and noise. If ℓ_g is the valid-supervised-token mean cross-entropy of view g, scale training minimizes $q^{-1}\sum_g \ell_g$, with $q = 4$. Identity continuation presents the same original-scale observations four times, matching families, presentations, and updates. A smooth group-loss control is retained in Appendix AB.

4 Experimental design

Families and exposure. The shared grammar covers additive, polynomial, bounded-denominator rational, periodic, exponential/logarithmic, and mixed-product relationships with one to four variables and 8–40 pre-serialization nodes. We generate 4,096 training and 1,024 development families, using 128 visible points from $[-1, 1]^d$ and relative Gaussian noise in $\{0, 0.01, 0.1\}$. A structural audit excludes coefficient-only and variable-renamed duplicates, the recorded scientific catalogue, and prior project exposures. Here, disjointness is defined by these recorded structural exclusions.

Each adapted arm takes 2,000 AdamW updates from shared pretrained weights, with 16 view presentations per update and a fixed final checkpoint. Original cohort A contains 256 families disjoint from training and development under the recorded structural key. Three noise levels and two coordinate views give 1,536 conditions per model; we evaluate both capacities and interfaces with three training seeds. Constant and capacity follow-ups reuse A, while the initialization controls use 16 exposed development families and six conditions each. Full model and return counts are recorded in the appendices.

Independent interface confirmation. Cohort B contains 256 newly generated families, structurally disjoint from recorded training, development, and previous evaluation families under the same exclusion rules. The four target–output configurations and primary high-accuracy interaction were fixed before its generation and evaluation. Six conditions per family and three existing 20M training seeds give 18,432 primary returns. The balanced-format checkpoints add 9,216 returns as a separately specified secondary comparison. Both cohorts share the supported grammar and scale range. Appendix A gives the complete design and seedwise results.

Returns and uncertainty. We use 64 proposals per view for synthetic queries and 1,024 for catalogue queries, charging both source budgets when searches are paired. Synthetic formulas are evaluated on 512 separate clean points; the appendices specify physical sampling. For a fixed returned

formula f, normalized risk is

$$R(f) = \frac{\sum_{i=1}^{N}(f(x_i) - f_\star(x_i))^2}{\sum_{i=1}^{N}(f_\star(x_i) - \bar{f}_\star)^2}. \tag{3}$$

We record high accuracy when $R^2 = 1 - R \geq 0.99$ and strict numerical success when $R \leq 2^{-23}$, evaluating algebraic equality separately. Any nonfinite prediction receives infinite risk and stays in every denominator and complete mean. Failed selection returns the visible mean and counts as a failure; observational tasks use held-out measured targets.

Training comparisons use 10,000 crossed family–training-seed bootstrap draws. For catalogue selection we resample 120 families, and for initialization we resample 16. Both 20M interface cohorts adjust nine comparisons with Bonferroni intervals: three endpoints, each with two simple effects and their interaction. Cohort B's twelve balanced-format contrasts form a separate secondary family; other intervals are pointwise. The bootstrap characterizes the observed seeds and families; subsequent comparisons on exposed families are exploratory. We record generation, refinement, and selection work together with service times measured during concurrent execution.

External prediction and physical recovery. A separate physical panel contains 24 previously exposed noisy conditions across nine SRSD laws and three ODE right-hand sides, with 1,024 proposals per model. The public slice has 24 SRBench-2025 datasets and three official splits (Imai Aldeia et al., 2025), fixed neural proposal counts, and recorded native budgets for other methods. The complete scale-trained public matrix crosses literal, role, and balanced targets with both outputs and three training seeds. Concrete and Airfoil also use three group-disjoint outer folds.

5 Results

5.1 An interface crossover confirmed on new families

Independent evaluation on cohort B confirms the ranking reversal in all three training seeds (Figure 2a; Table 2). Roles improve placeholder coverage over literals by $+3.86$ percentage points (95% interval $[2.21, 5.60]$); under constant requests, the roles-minus-literal difference is -2.37 ($[-4.23, -0.63]$; its adjusted interval includes zero). The prespecified interaction is -6.23 points, with nine-comparison adjusted interval $[-9.84, -2.79]$. Seedwise interactions are -6.77, -6.64, and -5.27. At $R^2 \geq .999$, the interaction is -7.44 points (adjusted $[-11.05, -3.86]$); strict numerical success gives -1.15 (nominal $[-2.45, 0.07]$).

Table 2: **Independent confirmation and balanced-format training on cohort B.** Each row pools 4,608 returns from three seeds. F/C denote placeholder/constant emission. Strict uses $R \leq 2^{-23}$; NF counts nonfinite conditions. Every complete mean FVU is infinite. Appendix A reports all seeds and contrast intervals.

Target	Output	$R^2 \geq .99$	$\geq .999$	Strict	Median FVU	NF
Literal	F	2,829 (61.39%)	1,597	416	0.00372	1
Literal	C	3,088 (67.01%)	1,908	450	0.00212	2
Roles	F	3,007 (65.26%)	1,822	428	0.00249	2
Roles	C	2,979 (64.65%)	1,790	409	0.00264	2
50/50	F	2,975 (64.56%)	1,808	420	0.00269	2
50/50	C	3,120 (67.71%)	1,922	441	0.00218	1

Exploratory paired comparisons on cohort A show that the interaction varies with continuation exposure. At 120M, identity continuation gives $+0.63$ points and scale training -5.88. The scale-minus-identity interaction difference is -6.51 points (pointwise 95% interval $[-9.44, -3.56]$) at $R^2 \geq .99$ and -10.39 ($[-13.43, -7.47]$) at .999 (Figure 5b; Appendix AD). Cohort A also gives -5.21 at 20M. Two subsequent seeds on exposed B families give -5.92 and -3.06 points (pooled -4.49, adjusted $[-9.00, -0.13]$), preserving both interface rankings across all five seeds (Appendix B).

5.2 BALANCED FORMAT EXPOSURE SUPPORTS BOTH OUTPUTS

Balanced training gives 2,975/4,608 high-accuracy placeholder returns and 3,120/4,608 constant returns. It improves literal/placeholder by $+3.17$ points (twelve-comparison adjusted interval $[0.32, 5.99]$) and role/constant by $+3.06$ ($[0.50, 5.82]$), with gains in every seed. Against the stronger specialists, observed differences are -0.69 points for placeholders (95% interval $[-1.93, 0.52]$) and $+0.69$ for constants ($[-0.65, 1.97]$; Figure 2b). Strict counts are 420/441 versus the specialists' 428/450. Table 2 retains all nonfinite outcomes. At 120M on exposed cohort A, balanced coverage remains close to both specialists and improves role/constant by $+4.32$ points (adjusted $[1.89, 7.04]$; Appendix M).

On a separate exposed 576-condition panel, repeated F and repeated C searches have 14 and 42 exclusive successes, respectively. At 64 total proposals, combining F+C gives 385 high-accuracy returns versus 400 for repeated C; the paired difference interval includes zero (Appendix S).

What reaches the numerical solver. Requests condition generation, and the emitted coefficient slots determine the fitting route. We trace 1,020,878 valid unique parents from the four cohort-B arms and another 510,010 from balanced models. Under constant requests, role training produces complete finite coefficient vectors in 76.68% of parents, compared with literal training's 89.24%; 11.55% of role parents contain placeholders. Balanced training produces 89.63% complete vectors and 0.02% placeholder parents (Figure 3a). Format exposure thus changes the numerical inputs available to refinement.

Direct request effects at coefficient entry. Holding observations and teacher histories fixed across 1,152 prefix pairs (16 families, nine checkpoints), adding `<mask_fittable>` shifts mean placeholder probability at FREE entry by 0.563–0.859; TYPED shifts are at most 1.31×10^{-7} (Figure 3b). This locates an immediate request effect before generated tokens change downstream context (Appendix F).

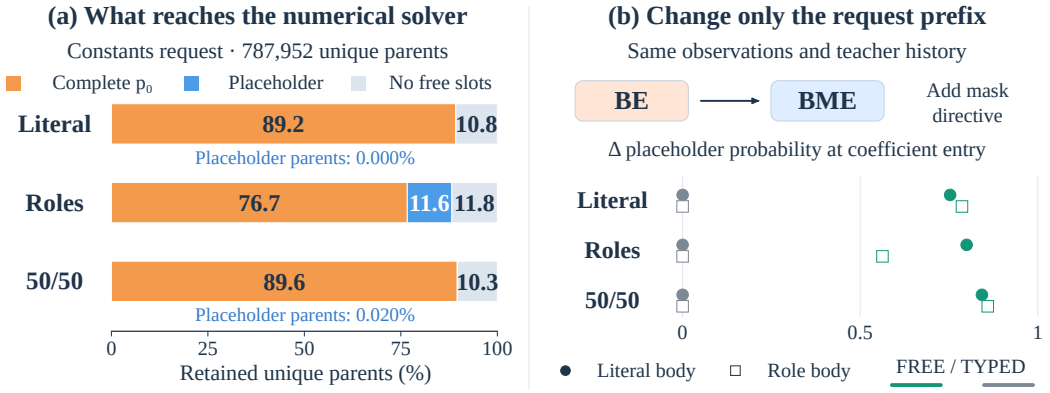


Figure 3: **Format exposure changes solver inputs and coefficient-entry probabilities.** (a) Under constants requests, three coefficient classes partition 787,952 retained unique parents. Labels show percentages. (b) BE is `<bos><expression>`; BME adds `<mask_fittable>`. The intervention holds observations and teacher histories fixed. Across 1,152 pairs, mean placeholder-probability changes at FREE entry are 0.563–0.859; TYPED changes are at most 1.31×10^{-7}. Markers distinguish literal and role teacher bodies.

5.3 NUMERICAL VALUES CONTRIBUTE AT DIFFERENT PIPELINE STAGES

Boundary supervision and numerical context. On 128 development families, zeroing CE weights on complete ten-token FREE spans reduces high-accuracy counts from 454/500 to 228/171 (placeholder/constant; Figure 4a). Zeroing only the eight value-byte losses retains 457/494; its paired difference intervals include zero. The boundary supervision and teacher-forced numerical history remain available in this control. On 32 families, coarsening or permuting FREE values in

that history raises downstream structural CE from 0.490 to 0.666/0.620 nats. These interventions distinguish direct value-byte supervision from the numerical context used to predict later tokens (Appendices AG.1 and AG).

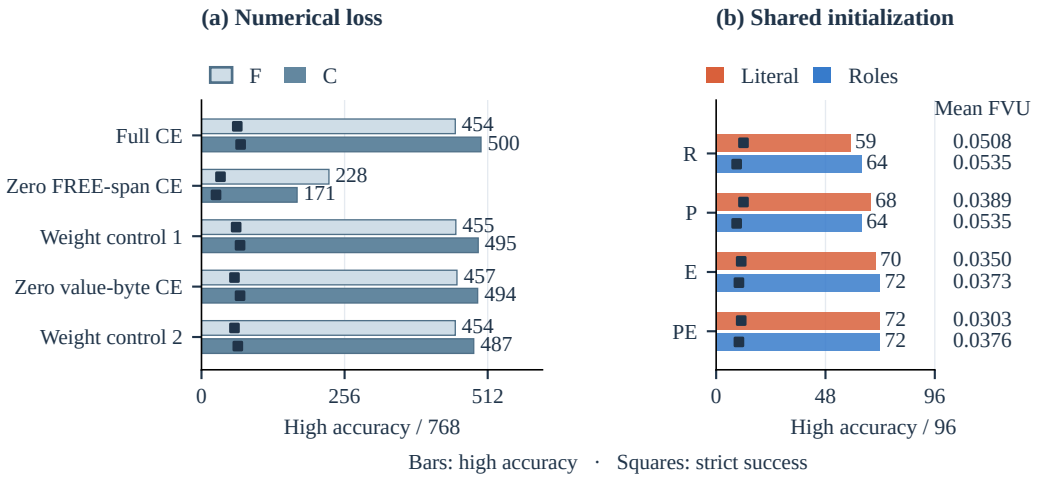


Figure 4: **Numerical supervision and transferable initialization.** (a) Loss controls on 768 conditions per output. Weight controls match the supervised mass of FREE-span or value-byte CE zeroing; teacher tokens are retained. (b) Fixed Literal/C and Roles/C parent pools on 96 conditions use eight starts: Gaussian (R), own prediction plus Gaussian (P), other expressions' empirical predictions (E), or empirical starts with the own prediction (PE). Bars show high accuracy; squares mark strict successes. Complete mean FVU appears at right. Appendix AE.2 gives paired intervals, bank construction, and work.

Initialization information transfers across candidate expressions. On 96 conditions from 16 exposed development families (one seed), fixed Literal/C pools give 59, 68, 70, and 72 high-accuracy returns under R/P/E/PE; Roles/C pools give 64, 64, 72, and 72 (Figure 4b). Empirical starts draw from other expressions' predictions, primarily for the same observation query. For Literal/C, E improves over R by $+11.46$ points (adjusted $[2.08, 22.92]$). Thus, useful starting values transfer across candidate expressions under shared fitting limits. Appendix AE.2 reports all precision endpoints, risks, paired intervals, and donor sources. Coarse and shuffled controls remain in Appendix J.2. A three-seed E2E continuation also supports informative initialization: predicted starts raise literal-trained $R^2 \geq .999$ returns from 89 to 126/288 ($+12.85$ points, adjusted $[4.17, 23.26]$; Appendix N).

A separate replay applies the same random-start fitting to all four configuration pools on 32 exposed conditions. Within each target format, placeholder and constant coverage coincide under four random starts in eight-parent pools; predicted-start replacement changes only literal/constant coverage (Appendix AE.1). This replay measures the contribution of refinement on fixed candidate pools; its interaction interval spans both positive and negative effects. Repeating this replay with trust-region reflective fitting preserves the four cold counts (Appendix K). Constant emission also uses more expression calls on a separate 32-family panel; the full accounting is in Appendix G.

Constrained emission in a second backbone. On the predesignated 48 exposed families, E2E literal/C and skeleton/P produce 511 and 370 high-accuracy returns out of 864; the mismatched cells produce none. The empty-generation counts identify failures before numerical fitting. The interaction persists with shared random initialization. These generation-time constraints complement the E2E initialization study (Appendix O).

The crossover across numerical precision. On the frozen cohort-B returns, a descriptive precision scan gives interactions of -4.36 points at $R^2 \geq .9999$ (pointwise interval $[-6.64, -2.15]$) and

-3.17 at $R^2 \geq .99999$ ($[-4.86, -1.48]$). Figure 5 shows this threshold dependence alongside the paired training-regime contrast; the three prespecified adjusted comparisons remain in Figure 2c.

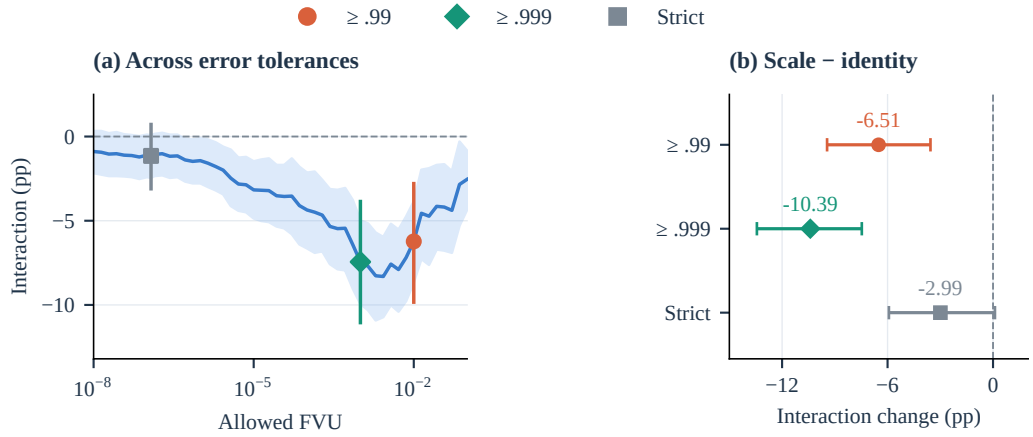


Figure 5: **The interaction across precision thresholds and training regimes.** (a) Continuous interaction on frozen cohort-B returns; shading is a pointwise 95% interval. Markers show the three prespecified endpoints with nine-comparison adjusted intervals. (b) Scale-minus-identity change in the supervision–interface interaction at 120M on exposed cohort A. Bars show pointwise 95% intervals from the paired exploratory contrast, including the strict endpoint (Appendix AD).

5.4 Public transfer evaluation

Public predictive performance. The complete public matrix crosses three targets with both outputs and three training seeds on 24 mixed real/synthetic datasets. Balanced training reaches 17/19/17 high-accuracy returns per 72 splits with placeholders and 19/17/17 with constants: 53/216 for each output. Every balanced seed exceeds the observed counts of native-budget Operon (12/72), PySR (10/72), and TIR (7/72). Paired dataset–seed intervals against PySR and Operon include zero, as does the public interaction interval: -0.46 points (adjusted $[-15.28, 13.89]$). This evaluation establishes the public coverage of both requests, with uncertainty about the direction of their interaction. Median FVU also favors different configurations: role/placeholder achieves 0.2500, balanced F/C 0.3014/0.2828, and Operon/TIR 0.2262/0.2268 (Figure 6). Each checkpoint retains 69 native returns and three mean fallbacks; all predictions are finite and strict counts are zero. Appendices T and I give all cells, intervals, and baseline configurations.

A separate 120M constants-request (C) follow-up uses 256 proposals, a 1,024-row cap, and RMS conditioning. Balanced checkpoints give 16/13/10 high-accuracy returns per 72 splits, versus frozen 120M's 9, PySR's 10, and Operon's 12 under native 120 s budgets. All adjusted high-accuracy difference intervals include zero; Appendix U reports all 432 returns and risks.

6 Discussion and conclusion

Crossed final-formula evaluation reveals a reproducible supervision–interface trade-off in scale-augmented Flash-ANSR continuation. Independent families confirm the ranking reversal at two prespecified precision thresholds; exploratory paired contrasts connect its magnitude to continuation exposure. The interventions distinguish boundary supervision, numerical history, and initialization, showing that useful starting values transfer across candidate expressions. Balanced 50/50 continuation provides a constructive response: one checkpoint improves both weaker pairings and reaches observed high-accuracy coverage within 0.70 points of the respective specialists. These findings support evaluating numerical supervision jointly with the requested output format and the solver's use of numerical predictions.

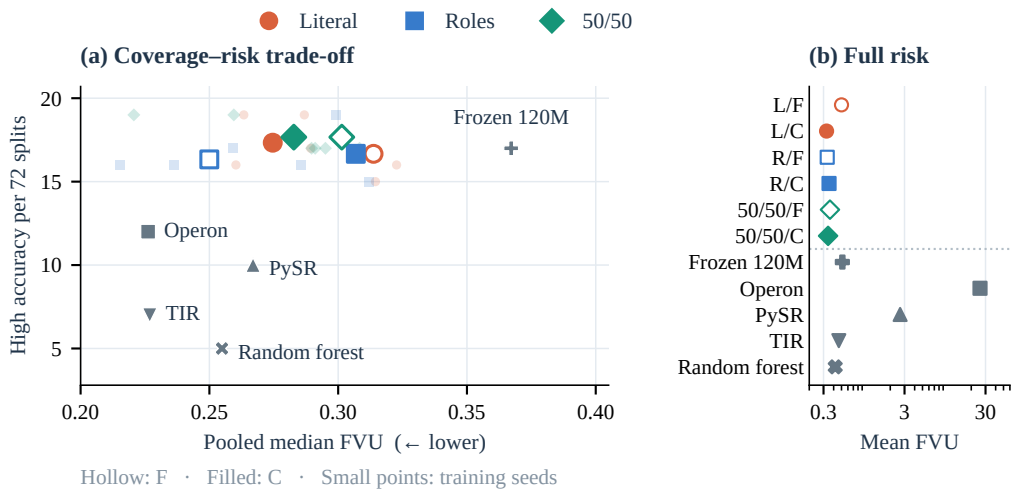


Figure 6: **Public coverage and complete prediction risk.** (a) Six public neural configurations: large markers show pooled high-accuracy counts per 72 splits and pooled median FVU; small markers show three seeds. (b) Complete mean FVU exposes a different risk ordering. L/R denote literal/role training. All fallbacks remain included. Neural queries use 1,024 proposals and all training rows; baselines retain native budgets. Paired cross-system intervals include zero (Appendix T).

Reproducibility Statement

Section 3 specifies the training and inference procedures. The appendices provide data-generation and exposure rules, optimization budgets, all-seed results, and complete counts, risks, and paired intervals. Tables and figures are generated from frozen returned-formula evaluations.

AI Use Statement

The authors used large language models solely to polish the text for clarity and assist with literature discovery. All cited references were manually verified against primary sources, and the authors take full responsibility for the paper's contents.

References

Deaglan J. Bartlett, Harry Desmond, and Pedro G. Ferreira. Exhaustive symbolic regression. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2211.11461.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 936–945, 2021. URL https://proceedings.mlr.press/v139/biggio21a.html.

Bogdan Burlacu, Gabriel Kronberger, and Michael Kommenda. Operon C++: An efficient genetic programming framework for symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1562–1570, 2020. doi: 10.1145/3377929.3398099.

Xieting Chu, Sriram Vishwanath, and Vijay Ganesh. Symbolic regression via latent iterative refinement. *arXiv preprint arXiv:2605.27245*, 2026. URL https://arxiv.org/abs/2605.27245.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL http://jmlr.org/papers/v25/23-0870.html.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023. URL https://arxiv.org/abs/2305.01582.

Fabricio Olivetti de Franca. Transformation-interaction-rational representation for symbolic regression. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2205.06807.

Jiao Hu, Jiaxu Cui, and Bo Yang. Learning interpretable network dynamics via universal neural symbolic regression. *Nature Communications*, 16:6226, 2025. doi: 10.1038/s41467-025-61575-7. URL https://www.nature.com/articles/s41467-025-61575-7.

Rodrigo Ibata, Wassim Tenachi, Foivos Diakogiannis, Neil Ibata, and Anirudh Shankar. NestyNet. III. symbolic regression from analytic neural surrogates. *arXiv preprint arXiv:2608.21051*, 2026. URL https://arxiv.org/abs/2608.21051.

Guilherme S. Imai Aldeia, Hengzhe Zhang, Geoffrey Bomarito, Miles Cranmer, Alcides Fonseca, Bogdan Burlacu, William G. La Cava, and Fabrício Olivetti de França. Call for action: Towards the next generation of symbolic regression benchmark. In *Genetic and Evolutionary Computation Conference Companion*, 2025. doi: 10.1145/3712255.3734309. URL https://arxiv.org/abs/2505.03977.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper/2022/hash/42eb37cdbefd7abae0835f4b67548c39-Abstract-Conference.html.

Gabriel Kronberger, Fabricio Olivetti de Franca, Deaglan J. Bartlett, Harry Desmond, and Pedro G. Ferreira. Guiding multi-objective genetic programming with description length improves symbolic regression solutions. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2605.22374.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabricio de Franca, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason Moore. Contemporary symbolic regression methods and their relative performance. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c0c7c76d30bd3dcaefc96f40275bdc0a-Abstract-round1.html.

Mikel Landajuela, Chak Shing Lee, Jiachen Yang, Ruben Glatt, Claudio P. Santiago, Ignacio Aravena, Terrell Mundhenk, Garrett Mulcahy, and Brenden K. Petersen. A unified framework for deep symbolic regression. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dbca58f35bddc6e4003b2dd80e42f838-Abstract-Conference.html.

Qian Li, Yuxiao Hu, Juncheng Liu, and Yuntian Chen. GenSR: Symbolic regression based on equation generative space. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8fd1a81c882cd45f64958da6284f4a3f-Abstract-Conference.html.

Yanjie Li, Jingyi Liu, Min Wu, Lina Yu, Weijun Li, Xin Ning, Wenqiang Li, Meilan Hao, Yusong Deng, and Shu Wei. MMSR: Symbolic regression is a multi-modal information fusion task. *Information Fusion*, 114:102681, 2025. doi: 10.1016/j.inffus.2024.102681. URL https://www.sciencedirect.com/science/article/pii/S1566253524004597.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. Rethinking symbolic regression datasets and benchmarks for scientific discovery. *Journal of Data-centric Machine Learning Research*, 2024. URL https://openreview.net/forum?id=qrUdrXsiXX.

Trent McConaghy. FFX: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, pp. 235–260. Springer, 2011. doi: 10.1007/978-1-4614-1770-5_13. URL https://trent.st/content/2011-GPTP-FFX-paper.pdf.

Kazem Meidani, Parshin Shojaee, Chandan K. Reddy, and Amir Barati Farimani. SNIP: Bridging mathematical symbolic and numeric realms with unified pre-training. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KZSEgJGPxu.

Vikas Palakonda, Samira Ghorbanpour, Sangseok Yun, Il-Min Kim, Jae-Mo Kang, and Sunghwan Moon. A comprehensive survey on symbolic regression: State-of-the-art approaches, key applications, benchmark evaluations, and future research directions. *Archives of Computational Methods in Engineering*, 2026. doi: 10.1007/s11831-026-10681-w. URL https://link.springer.com/article/10.1007/s11831-026-10681-w.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Paul Saegert and Ullrich Köthe. Breaking the simplification bottleneck in amortized neural symbolic regression. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64015.

Philipp Scholl, Katharina Bieker, Hillary Hauger, and Gitta Kutyniok. ParFam—(neural guided) symbolic regression via continuous global optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8y5Uf6oEiB.

Parshin Shojaee, Kazem Meidani, Amir Barati Farimani, and Chandan K. Reddy. Transformer-based planning for symbolic regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ffb4e3118280a66b192b6f06e0e2596-Abstract-Conference.html.

Wassim Tenachi, Rodrigo Ibata, and Foivos I. Diakogiannis. Deep symbolic regression for physics guided by units constraints: Toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2):99, 2023. doi: 10.3847/1538-4357/ad014c. URL https://arxiv.org/abs/2303.03192.

Martin Vastl, Jonáš Kulhánek, Jiří Kubalík, Erik Derner, and Robert Babuška. SymFormer: End-to-end symbolic regression using transformer-based architecture. *IEEE Access*, 12:37840–37849, 2024. doi: 10.1109/ACCESS.2024.3374649. URL https://erik-derner.github.io/research/files/vastl2024symformer.pdf.

Henrik Voigt, Paul Kahlmeyer, Kai Lawonn, Michael Habeck, and Joachim Giesen. Analyzing generalization in pre-trained symbolic regression. *arXiv preprint arXiv:2509.19849*, 2025. URL https://arxiv.org/abs/2509.19849.

Shengbin Ye and Meng Li. Ab initio nonparametric variable selection for scalable symbolic regression with large p. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72041–72062, 2025. URL https://proceedings.mlr.press/v267/ye25f.html.

Zihan Yu, Jingtao Ding, Yong Li, and Depeng Jin. Symbolic regression via MDL-former guided search: From minimizing prediction error to minimizing description length. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a402493de088886740b5939f666a6e56-Abstract-Conference.html.

Hengzhe Zhang, Qi Chen, Bing Xue, Wolfgang Banzhaf, and Mengjie Zhang. RAG-SR: Retrieval-augmented generation for neural symbolic regression. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/19a8e70828c01059631f913442ae31e6-Abstract-Conference.html.

Wenxiao Zhao, Dong Liu, Kaiyi Xu, Feng Liu, Zhen Zhao, Fei Ben, Shu Wang, Wenhao Li, Ying Nian Wu, Fenghua Ling, Haobo Li, and Lei Bai. A-SR: Self-evolving agentic LLMs for symbolic regression via hierarchical coordination. *arXiv preprint arXiv:2608.04872*, 2026. URL https://arxiv.org/abs/2608.04872.

Peiyi Zheng, Yanming Kang, Hans De Sterck, and Giang Tran. MOSAIC-SR: Transformer-guided symbolic regression for scientific equation recovery. *arXiv preprint arXiv:2609.20997*, 2026. URL https://arxiv.org/abs/2609.20997.

GUIDE TO THE SUPPLEMENTARY EVIDENCE

The main paper follows the interface crossover, balanced training, and interventions on the generation–refinement pathway. The supplementary material provides the complete measurements and the following routes through the evidence. Figure 7 summarizes the numerical workflow.

Reader's question	Evidence to consult
Does the ranking reversal replicate?	Independent confirmation (§A), later seeds (§B), capacity (§AD), and continuous precision (§L).
What does one shared checkpoint retain?	Balanced comparisons (§A; Figure 24), 120M follow-up (§M), and mixed-query deployment (§S).
Where do numerical values act?	Fixed-history requests (§F), actual coefficient routes (§G), loss and context (§AG), initialization and empirical coefficient distributions (§AE; §AE.2), E2E initialization (§N) and constrained emission (§O).
Which conditions affect the result?	Structural motifs and budgets (§J), token weighting (§Q), alternative fitting (§K), and scale conditions (§AD; §H).
How do the public returns compare?	Complete two-interface matrix (§T), native baseline settings (§I), and the separate 120M short-budget experiment (§U).
What do coordinates and physical recovery add?	Coordinate study (§AK), RMS transfer (§V), physical recovery (§AH), and complete risk tails (§X).

Each study specifies its sampling unit, exposure status, full denominator, and interval family. Historical seed-0 results and the current three-seed matrix are labeled separately. Extended coordinate and physical studies include their successful returns, normalization controls, and risk outcomes together.

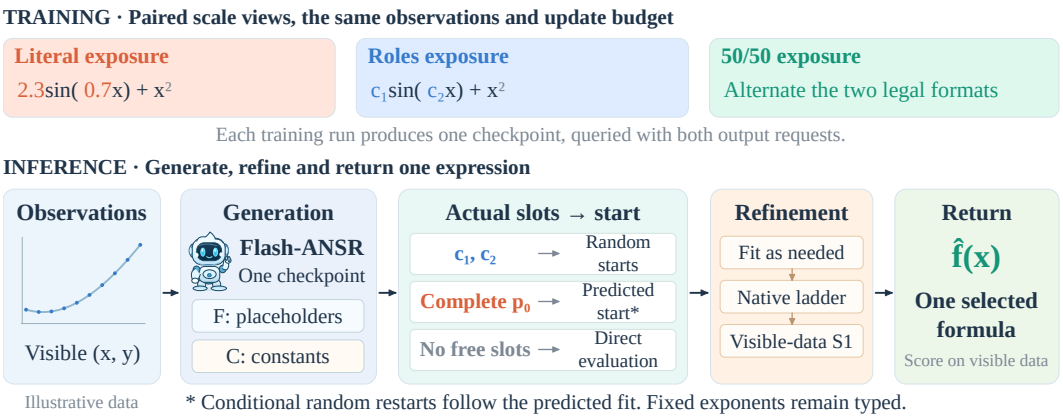


Figure 7: **From observations and output requests to a returned formula.** The generated coefficient slots select random initialization, predicted initialization, or direct evaluation. Numerical fitting, discrete refinement, and visible-data S1 selection complete the return. Each training strategy produces one checkpoint per seed; the curve and formulas are illustrative.

A INDEPENDENT INTERFACE CONFIRMATION AND BALANCED-FORMAT TRAINING

Cohorts and analysis order. Cohort A is the original 256-family study: its constant-output and capacity comparisons followed earlier placeholder readout. Cohort B was generated for the interface interaction after those observations. Its six fixed 20M scale-trained checkpoints, both requested outputs, all six family conditions, and the primary $R^2 \geq .99$ interaction were specified before data generation and clean evaluation. Structural exclusion uses the same coefficient- and variable-renaming rules against recorded training, development, and prior evaluation families. The exclusion reference comprises the recorded project families and catalogues; overlap with the full upstream pretraining corpus is unassessed. The grammar and scale support match the original study.

The complete primary comparison has $256 \times 6 \times 3 \times 4 = 18,432$ returns. The secondary 50/50 control has three additional fixed final checkpoints, each evaluated through both outputs, adding 9,216 returns. All models and final checkpoints are fixed before evaluating cohort B.

Balanced continuation. Training preserves the original pretrained initialization, paired observations, scale views, 2,000 updates, view presentations, and valid-token mean loss. For each four-view family presentation, two views are chosen uniformly for role targets and the remaining two use literals, using a separate precomputed random schedule. Each run therefore presents 16,000 examples in each legal native instruction/target format. The resulting checkpoint serves both output directives within a seed. This control changes supervision composition and format exposure together.

Endpoints and intervals. All 27,648 returns are evaluated on 512 separate clean points per condition using the risk in Equation 3. Nonfinite and empty outcomes remain in denominators. The bootstrap crosses 256 resampled families with three globally resampled training seeds for 10,000 replicates; all six conditions of a family stay together. Primary Bonferroni intervals use nine comparisons: the two role-minus-literal simple effects and their interaction at each of three thresholds. Secondary intervals adjust twelve comparisons separately: 50/50 minus each specialist under each output at three thresholds. Resampling characterizes variability among the three observed training seeds and evaluation families.

Table 3: Cohort B seedwise counts. Triplets are seeds 0/1/2, each with 1,536 conditions. F/C denote placeholder/constant output. NF counts conditions with at least one nonfinite evaluation prediction.

Target	Output	$R^2 \geq .99$	$R^2 \geq .999$	Strict	NF
Literal	F	938 / 935 / 956	532 / 532 / 533	138 / 134 / 144	0 / 1 / 0
Literal	C	1031 / 1027 / 1030	636 / 637 / 635	152 / 148 / 150	1 / 0 / 1
Roles	F	996 / 1006 / 1005	594 / 616 / 612	137 / 143 / 148	0 / 1 / 1
Roles	C	985 / 996 / 998	592 / 600 / 598	140 / 132 / 137	1 / 1 / 0
50/50	F	989 / 1000 / 986	601 / 605 / 602	150 / 143 / 127	1 / 1 / 0
50/50	C	1046 / 1029 / 1045	646 / 623 / 653	149 / 146 / 146	0 / 0 / 1

Coverage and returned-formula risk. Every pooled target/output cell contains one or two nonfinite conditions, so every complete mean FVU is infinite. Table 2 reports full denominators, median risk, and both high-precision and strict counts alongside coverage. The balanced models improve the weaker single-format combination for both outputs; differences from the stronger specialists remain unresolved.

Table 4: Cohort B paired differences and percentile intervals in percentage points. The upper block adjusts nine primary comparisons; the lower block adjusts twelve secondary comparisons. Strict means $R \leq 2^{-23}$.

Contrast	Endpoint	Difference (pp)	95% interval	Adjusted interval
Roles−literal, F	$R^2 \geq .99$	+3.86	[2.21, 5.60]	[1.50, 6.36]
Roles−literal, C	$R^2 \geq .99$	-2.37	[-4.23, -0.63]	[-5.16, 0.13]
Interaction	$R^2 \geq .99$	-6.23	[-8.66, -3.86]	[-9.84, -2.79]
Roles−literal, F	$R^2 \geq .999$	+4.88	[3.04, 6.73]	[2.15, 7.55]
Roles−literal, C	$R^2 \geq .999$	-2.56	[-4.30, -0.80]	[-5.14, 0.00]
Interaction	$R^2 \geq .999$	-7.44	[-9.94, -4.99]	[-11.05, -3.86]
Roles−literal, F	Strict	+0.26	[-0.59, 1.17]	[-1.02, 1.55]
Roles−literal, C	Strict	-0.89	[-1.93, 0.15]	[-2.41, 0.65]
Interaction	Strict	-1.15	[-2.45, 0.07]	[-3.10, 0.72]
50/50−literal, F	$R^2 \geq .99$	+3.17	[1.30, 5.12]	[0.32, 5.99]
50/50−literal, C	$R^2 \geq .99$	+0.69	[-0.65, 1.97]	[-1.37, 2.60]
50/50−roles, F	$R^2 \geq .99$	-0.69	[-1.93, 0.52]	[-2.60, 1.17]
50/50−roles, C	$R^2 \geq .99$	+3.06	[1.30, 4.82]	[0.50, 5.82]
50/50−literal, F	$R^2 \geq .999$	+4.58	[2.91, 6.25]	[1.95, 7.06]
50/50−literal, C	$R^2 \geq .999$	+0.30	[-1.52, 2.04]	[-2.48, 2.86]
50/50−roles, F	$R^2 \geq .999$	-0.30	[-1.82, 1.17]	[-2.59, 1.87]
50/50−roles, C	$R^2 \geq .999$	+2.86	[0.72, 4.95]	[-0.41, 5.94]
50/50−literal, F	Strict	+0.09	[-1.24, 1.37]	[-2.02, 1.96]
50/50−literal, C	Strict	-0.20	[-1.04, 0.65]	[-1.50, 1.11]
50/50−roles, F	Strict	-0.17	[-1.52, 1.11]	[-2.34, 1.69]
50/50−roles, C	Strict	+0.69	[-0.26, 1.65]	[-0.72, 2.18]

Table 5: Original cohort A scale-trained returns at both capacities, retained with all seeds. Each seed has 1,536 conditions. Complete means pool all seeds and retain infinite risk.

Size	Target	Emission	$R^2 \geq .99$	Strict	NF	Mean FVU
20M	Literal	F	944 / 949 / 949	167 / 166 / 162	1 / 0 / 0	∞
20M	Roles	F	993 / 996 / 1007	188 / 169 / 170	0 / 0 / 1	∞
20M	Literal	C	1037 / 1037 / 1040	179 / 167 / 186	0 / 0 / 1	∞
20M	Roles	C	1022 / 1003 / 1003	163 / 156 / 155	0 / 0 / 0	0.1285
120M	Literal	F	1133 / 1111 / 1145	276 / 295 / 281	0 / 0 / 0	0.02898
120M	Roles	F	1156 / 1155 / 1173	292 / 285 / 296	0 / 1 / 0	∞
120M	Literal	C	1251 / 1255 / 1249	300 / 298 / 313	0 / 1 / 0	∞
120M	Roles	C	1192 / 1186 / 1201	265 / 266 / 264	1 / 0 / 0	∞

B ADDITIONAL TRAINING-SEED FOLLOW-UP

We repeated the original literal, roles, and balanced mixed scale-ERM recipes with training seeds 20260924021 and 20260924121. All six last-step checkpoints were evaluated through both native output interfaces on all 256 B families and six conditions, adding 18,432 returns. These later training runs repeat the evaluation on exposed cohort-B families. The original independent confirmation comprises the first three seeds, with its original confidence intervals.

Table 6: Complete success counts, shown as $R^2 \geq 0.99$ / strict (FVU $\leq 2^{-23}$). F and C denote native placeholder and numeric emissions. Every column retains empty and nonfinite returns. The five-seed column is a follow-up aggregate.

Target	Output	Original 3 $n = 4608$	New s3 $n = 1536$	New s4 $n = 1536$	All 5 $n = 7680$
Literal	F	2829 / 416	935 / 134	972 / 143	4736 / 693
Literal	C	3088 / 450	1038 / 143	1036 / 152	5162 / 745
Roles	F	3007 / 428	995 / 146	1003 / 150	5005 / 724
Roles	C	2979 / 409	1007 / 142	1020 / 139	5006 / 690
Mixed	F	2975 / 420	987 / 135	1007 / 139	4969 / 694
Mixed	C	3120 / 441	1043 / 148	1047 / 144	5210 / 733

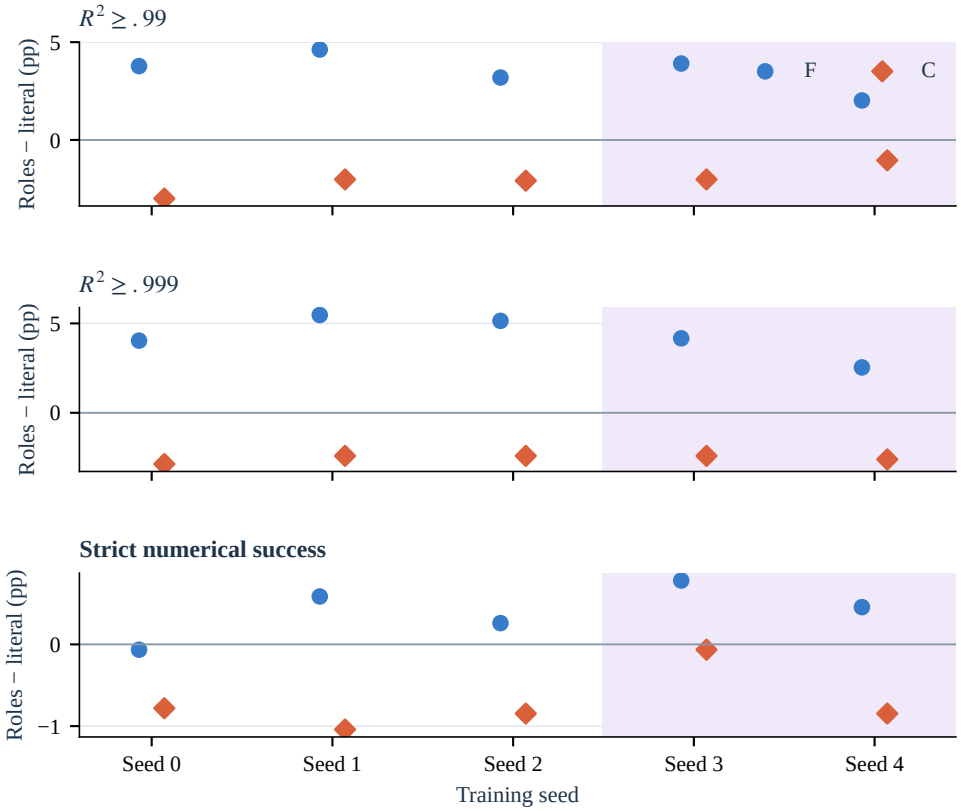


Figure 8: Per-seed differences between roles and literal training under each emission interface. Each dot is a training run's mean over all 1,536 conditions. Shading identifies the seeds added after the original confirmation.

The interaction $[(\text{roles} - \text{literal})_C - (\text{roles} - \text{literal})_F]$ in the two additional seeds is $R^2 \geq 0.99$: -4.49 [-9.00, -0.13]; $R^2 \geq 0.999$: -5.86 [-10.35, -1.56]; strict: -1.07 [-3.53, +1.19] percentage points (95% Bonferroni-adjusted intervals across the nine interface contrasts). The interaction $[(\text{roles} - \text{literal})_C - (\text{roles} - \text{literal})_F]$ in the five-seed follow-up is $R^2 \geq 0.99$: -5.53 [-8.98, -2.20]; $R^2 \geq 0.999$: -6.81 [-10.11, -3.54]; strict: -1.12 [-2.70, +0.29] percentage points (95% Bonferroni-adjusted intervals across the nine interface contrasts). For high accuracy in the five-seed follow-up, balanced training differs from literal/F: +3.03 [+0.74, +5.30]; literal/C: +0.62 [-1.11, +2.24]; roles/F: -0.47 [-2.12, +1.21]; roles/C: +2.66 [+0.48, +5.10] percentage points (95% intervals

adjusted across the 12 mixed-training contrasts). All four adjusted strict-success intervals include zero. In the new-two-only analysis, all 12 adjusted mixed-training intervals include zero. Across the five seeds, literal/F, literal/C, roles/F, roles/C, mixed/F, and mixed/C have 2, 3, 5, 3, 4, 1 nonfinite returns, respectively. Complete mean FVU is infinite in every cell; all nonfinite returns remain in the success denominators. All 12 mixed-training contrasts per cohort, all view/noise strata, nonfinite outcomes, and timing records are retained in the accompanying machine-readable results. Intervals use the fixed crossed family/training-seed bootstrap, with all six conditions of a family kept together.

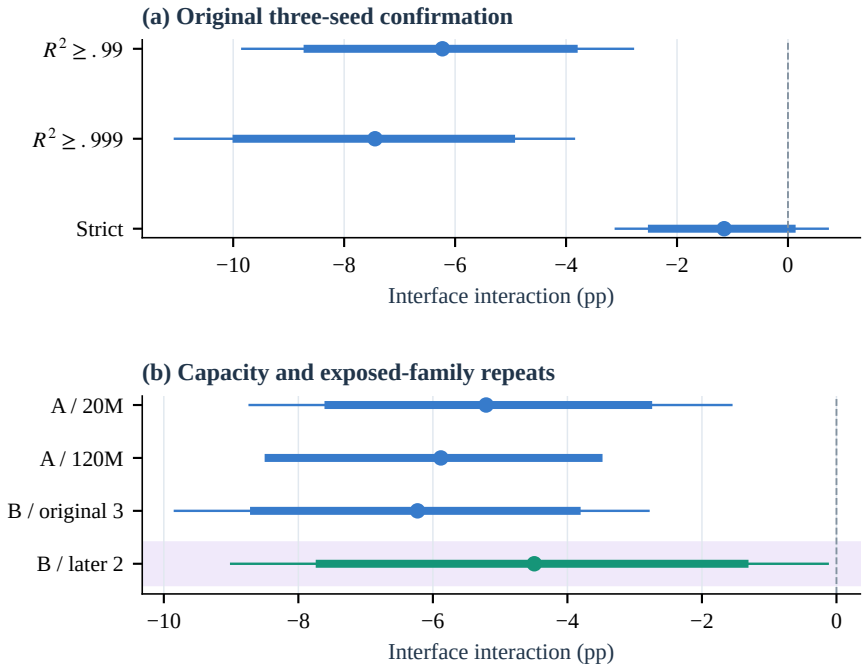


Figure 9: **Precision, capacity, and later training repeats.** (a) Original three-seed cohort-B confirmation: thick nominal 95% intervals and thin nine-comparison adjusted intervals. (b) High-accuracy interactions. The two capacities share cohort A. B / orig. 3 is the independent confirmation; B / new 2 reuses exposed B families for subsequent training seeds.

C PUBLIC TRAINING-SEED REPLICATION

The raw-coordinate placeholder comparison repeats literal/role and identity/scale arms with all three existing final 20M checkpoints on the same 24 datasets and three official splits. The two additional seeds add 576 returns to the original 288, with no checkpoint or dataset selection. Each supported query uses 1,024 proposals and the original native solver, refinement, and selector. The same unsupported-input policy retains three training-mean fallbacks per checkpoint. The initial role-scale seed-0 execution used two fitting workers for 42 supported returns; the added seeds use the canonical one-worker execution. Worker counts are shared; elapsed time is recorded separately for each run.

The role-scale counts are 16/16/17 versus identity's 13/15/14; median and mean FVU improve in each seed. Literal counts are 15/16/19 versus 13/15/13. Pooled role-scale median/mean FVU is 0.2500/0.3332; literal-scale is 0.3137/0.5002. The seed-0 table below remains a separate historical comparison including standardized-coordinate checkpoints and all external methods.

Table 7: All raw adapted public checkpoints. Triplets correspond to training seeds 0/1/2, each evaluated on all 72 splits. Pooled means retain 216 returns. Every checkpoint has 69 native returns, three fallbacks, zero strict successes, and zero nonfinite predictions.

Target	Training	High accuracy (0/1/2)	Median FVU (0/1/2)	Pooled mean
Literal	Identity	13 / 15 / 13	0.3556 / 0.3453 / 0.3579	0.8480
Literal	Scale	15 / 16 / 19	0.3145 / 0.3227 / 0.2869	0.5002
Roles	Identity	13 / 15 / 14	0.3209 / 0.2886 / 0.3546	0.7595
Roles	Scale	16 / 16 / 17	0.2364 / 0.2154 / 0.2592	0.3332

Table 8: Exploratory public high-accuracy differences. Bootstrap draws jointly resample 24 datasets and three training seeds, retaining all three official splits within each dataset. There are 10,000 draws. The last column adjusts four comparisons; fixed external baselines are shared across training seeds.

Role-scale minus	Difference (pp)	Nominal 95%	Adjusted 95%
roles identity	+3.24	$[-4.17, 11.11]$	$[-6.94, 13.89]$
PySR	+8.80	$[-4.17, 23.61]$	$[-6.94, 27.78]$
Operon	+6.02	$[-9.72, 22.69]$	$[-14.35, 27.78]$
frozen120	-0.93	$[-10.65, 6.94]$	$[-13.89, 9.72]$

These repeated training runs measure adaptation variability on the same previously exposed dataset slice. Recorded native budgets differ across systems; the full public competition uses 30 official splits and 3,600 seconds per run.

D VISUAL SUMMARIES AND FULL COMPARISON TABLES

The following displays summarize the same completed returns used in the main text. The paired intervals, precision tiers, and dataset matrix provide complementary views of those measurements.

Table 9: Full coordinate-policy comparison: 120 SRSD families, 720 conditions, 2,048 total proposals per row. Every complete mean FVU is infinite. $R > 10$ includes nonfinite cases; bits price the original-coordinate returned formula.

Pool / selection	$R^2 \geq .99$	$\geq .999$	Strict	NF	$R > 10$	Median bits
Raw + raw	517	489	260	5	24	50.3
RMS + RMS	547	508	217	5	22	364.5
Raw + RMS, native S1	585	541	284	5	15	185.9
Raw + RMS, error	585	542	252	6	17	120.7
Raw + RMS, transported S1	552	522	289	3	15	56.2

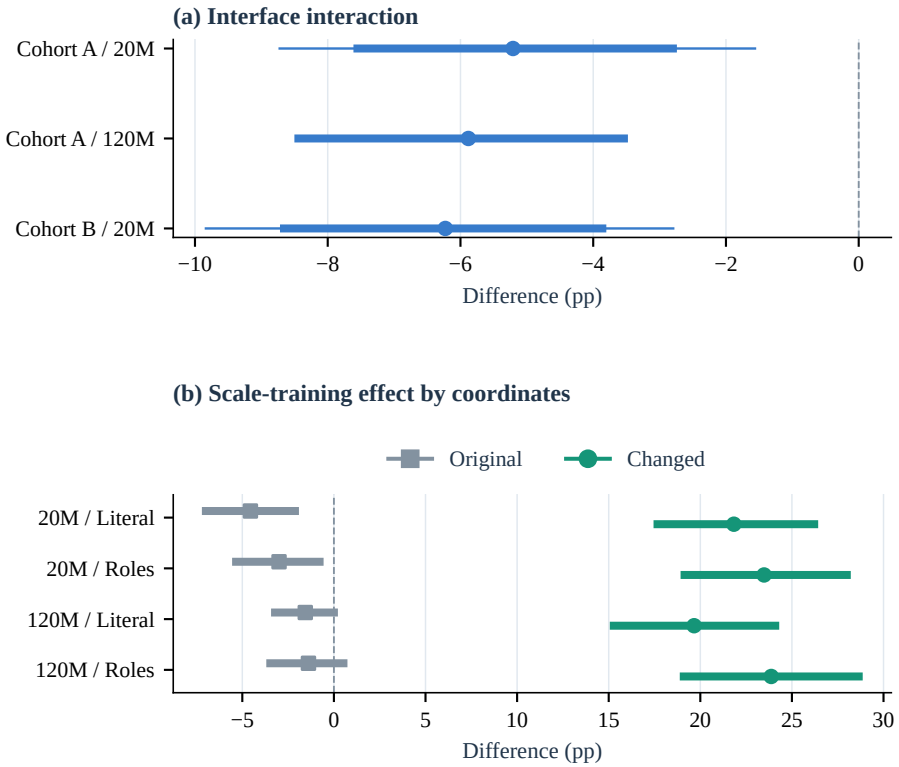


Figure 10: Interface interactions and scale-training effects, from the original and independent cohorts. The scale panel separates original and changed coordinates; intervals are paired, pointwise 95%.

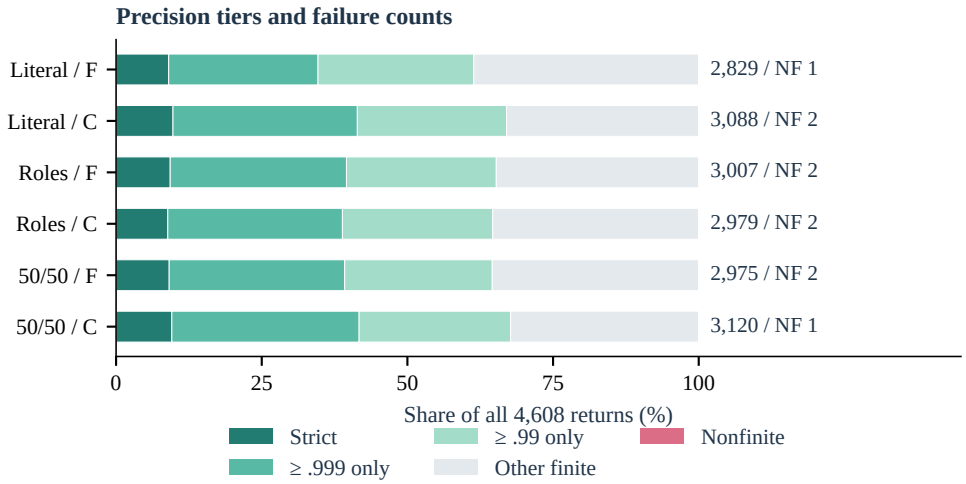


Figure 11: All six cohort B configurations. The precision segments are mutually exclusive. Strict means $R \leq 2^{-23}$; the next two bins add returns reaching $R^2 \geq .999$ and .99, respectively. Nonfinite counts are labeled separately because the bars are too small to distinguish at this scale.

Table 10: Public short-budget slice: 24 datasets, three official splits, and all 72 outcomes including labeled mean fallbacks. Neural methods use 1,024 proposals; other methods retain native budgets. Adapted models in this table use one checkpoint. Full configurations appear in Appendix AI.

System / coordinates	$R^2 \geq .99$	Median FVU	Mean FVU
Frozen 120M / raw	17/72	0.3672	0.5124
Frozen 20M / raw	17/72	0.4436	0.5525
Literal identity / raw	13/72	0.3556	0.6432
Literal scale / raw	15/72	0.3145	0.7931
Roles identity / raw	13/72	0.3209	0.8688
Roles scale / raw	16/72	0.2364	0.3588
Roles scale / standardized	4/72	0.3614	0.8503
PySR	10/72	0.2670	2.6482
Operon	12/72	0.2262	25.6476
TIR	7/72	0.2268	0.4633
Random forest	5/72	0.2549	0.4185

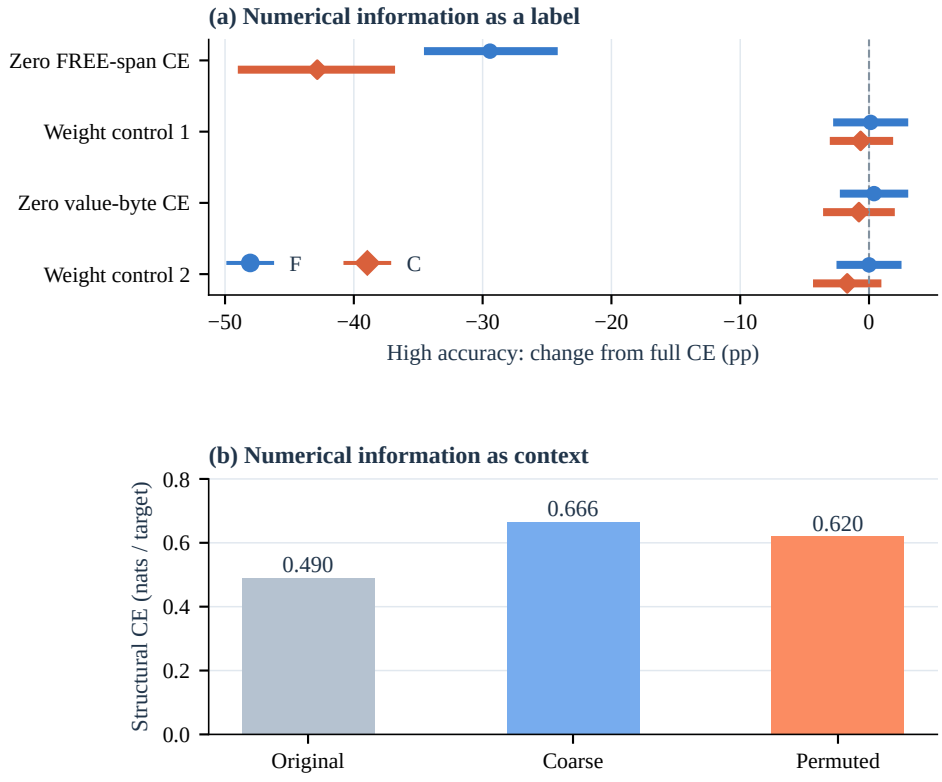


Figure 12: Numerical loss and teacher-forced context interventions. (a) High-accuracy changes from full CE on 768 conditions from 128 development families, with the original paired 95% intervals. Both weight controls preserve the corresponding supervised mass. (b) Structural CE on 32 families at the fixed literal-scale checkpoint; these are descriptive means. The two interventions probe numerical labels and teacher-provided context separately.

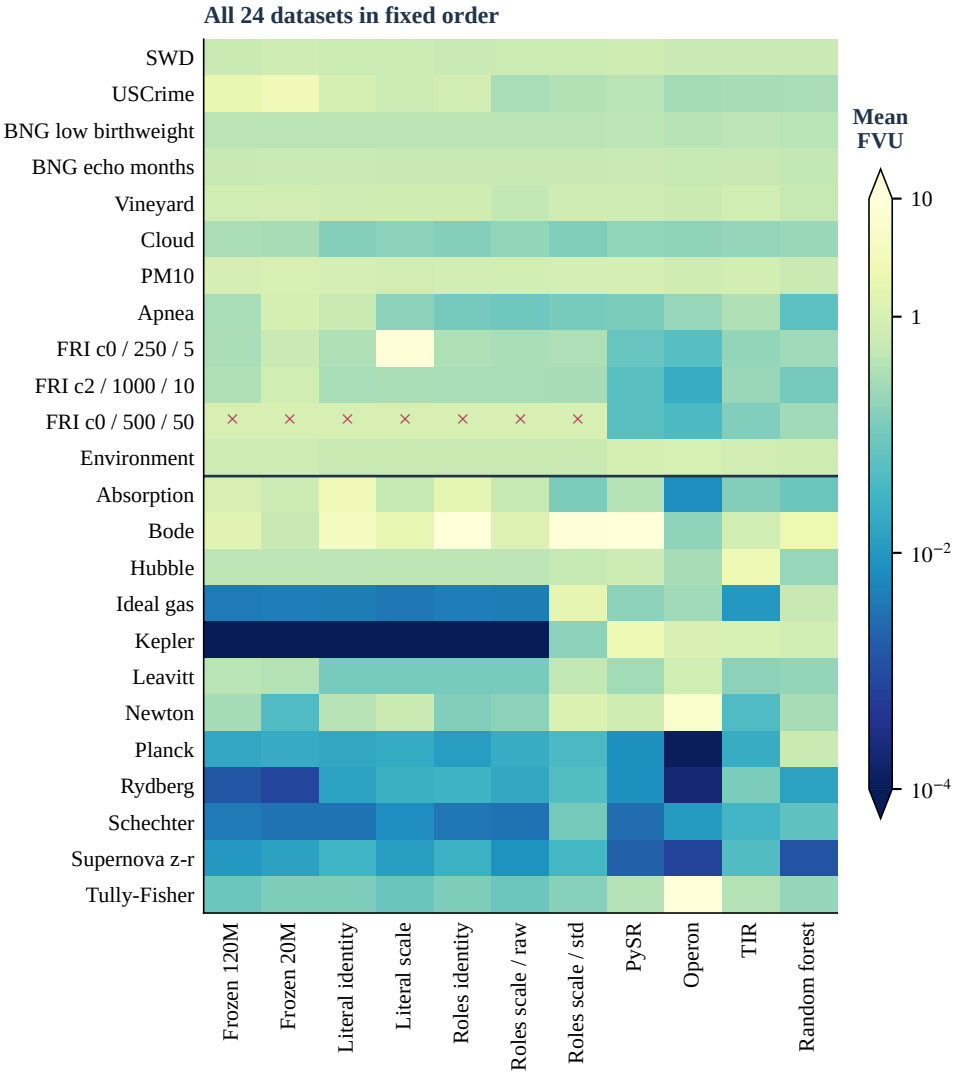


Figure 13: All 24 public datasets and the original seed-0 configurations of the 11 historical systems in the seed-0 public table. Each cell averages FVU over three official splits. Colors are clipped to $[10^{-4}, 10]$ for visibility; no value is omitted from the summaries. Crosses identify at least one documented non-native mean fallback. Dataset order follows the fixed source identifiers.

E COORDINATE TRANSPORT AND THE RETURNED DESCRIPTION

For the RMS transport in Equation 6, the same mathematical function has unchanged FVU in exact arithmetic. Description length is charged to the returned expression in its original coordinates: FVU is invariant for the same function under exact-arithmetic transport, but the native score charges the returned spelling:

$$S_x(\mathcal{T}_s h) - S_z(h) = \frac{2}{n \log_2 10}[L_x(\mathcal{T}_s h) - L_z(h)]. \tag{4}$$

Scale constants can be expensive, and discrete refinement need not commute with transport. All reported lengths price the original-coordinate formula.

We compare selection between saved winners by their stored native S1, visible FVU alone, or transported S1. Native/error rules use the same finite-score eligibility and error floor; ties prefer the first search. Raw+RMS uses raw as that first search. A repeated-coordinate control adds a second search with seed offset 1,000,003, with the same checkpoint, observations, and per-view budget. The old transported rule retains its original eligibility. Selection is frozen before clean evaluation. G2 separately admits certified original-domain executions, including documented removable extensions; failed admission remains in the full denominator (Appendix AF).

F DIRECT REQUEST-PREFIX EFFECTS UNDER FIXED HISTORIES

We held each observation and teacher-forced body fixed and changed BE to BME, adding the native `<mask_fittable>` directive. The first 16 existing development families by ID cover all six generator categories. Four saved scale views, two body formats, and nine checkpoints give 1,152 BE/BME pairs (2,304 sequences). The probe uses native inference's all-NaN numeric channel and raw vocabulary logits. Each body format retains its own canonicalization and receives a separate analysis. All saved observations and body mappings were independently reconstructed.

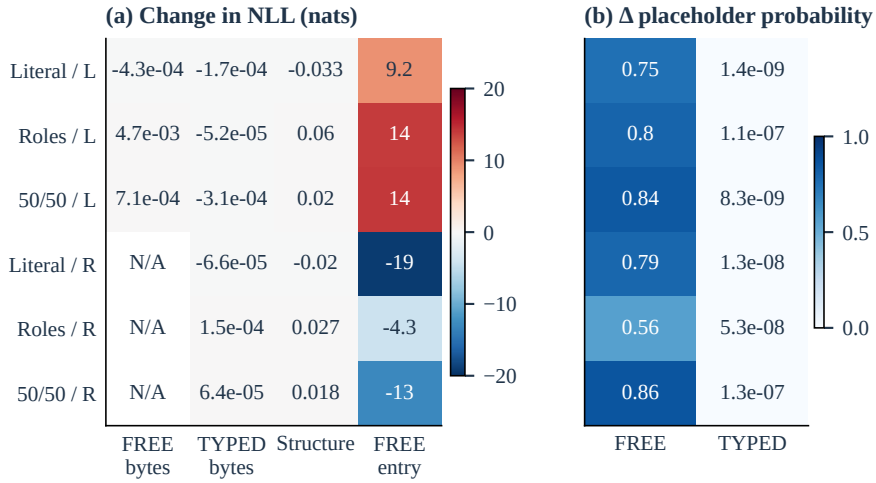


Figure 14: Fixed-history prefix effects, pooled over three checkpoints per training target. Panel A shows token-weighted NLL differences (BME minus BE), separating IEEE value bytes, structure, and the FREE emission start. N/A marks FREE value bytes absent from role bodies; its start token is the placeholder, while the literal body's start is the IEEE opener. Panel B shows the corresponding placeholder-start probability change, separately for FREE and TYPED sites. These are paired comparisons within one unchanged body, not between the two body formats.

The mean FREE-start probability shift is 0.563–0.859, whereas the mean TYPED shift is at most 1.31×10^{-7}. For literal bodies, the mean IEEE-start loss increases 9.19–14.09 nats, but the mean FREE-byte loss change is within 0.00475 nats. Mean TYPED-byte and structure changes are within 0.00031 and 0.0601 nats. The mean fixed-history effect concentrates on the emission-format decision, locating a direct request effect at coefficient entry.

G ACTUAL OUTPUT TYPES AND NUMERICAL WORK

The F and C requests condition generation while allowing both numerical token types. We decode all retained unique parents from the four cohort-B configurations and the two balanced configurations. Each row in Table 11 covers 4,608 complete returns. The four primary arms contain 1,020,878 parents in total; balanced F/C adds 242,908/267,102. All parents decode successfully. This archive follows native validity checks, simplification, and deduplication, so its denominator differs from the 64 raw proposals per query.

A parent is placed in one of three disjoint classes: at least one native placeholder; a complete finite vector for all free coefficients; or no free coefficient. The classification preserves typed numerical sites and shared parameter identities. A placeholder requests random fitting even when the user asked for constants; a complete vector enables the predicted-start path. The last two columns describe additional, overlapping properties. "Typed" means the parent has a fixed exponent, root index, or other typed literal. The "Refined typed" column counts parents with a retained descendant whose fixed typed constants were fitted. Its denominator is retained parents; attempted refinements are accounted for separately.

Table 11: **Actual parent types under requested interfaces.** Fractions use all retained unique parents in each row. Placeholder, complete p_0, and no-free columns sum to 100%; typed and thaw columns overlap these categories.

Target	Request	Parents	Placeholder (%)	p_0 (%)	No free (%)	Typed (%)	Refined typed (%)
Literal	F	257,417	86.136	0.017	13.848	53.545	20.616
Literal	C	266,122	0.000	89.241	10.759	54.212	18.874
Roles	F	242,611	89.551	0.003	10.446	55.750	19.838
Roles	C	254,728	11.554	76.675	11.771	58.007	21.594
50/50	F	242,908	89.377	0.004	10.619	54.046	19.109
50/50	C	267,102	0.020	89.634	10.346	55.805	19.290

Under C requests, literal training supplies complete coefficients in 89.24% of parents, role training in 76.68%, and balanced training in 89.63%. Role-trained C output contains placeholders in 11.55% of parents, while balanced C output reduces this fraction to 0.02%. These observations connect format exposure to the objects passed to refinement. The generated expression structures and coefficient distributions also vary across models, and jointly contribute to the full-system comparison.

Actual refinement calls. The completed scale-sensitivity follow-up supplies a separate instrumented four-arm panel: 32 fixed, already exposed cohort-B families, three noise levels, identity and scale-six views, and three training seeds, giving 576 queries per arm. All use 64 proposals and the same native numerical policy. Table 12 counts expression forwards, including numerical Jacobian probes, failed fits, scoring, and discrete refinement. These totals are broader than an optimizer's reported `nfev`. The proposals come from new follow-up queries instrumented to record every forward evaluation.

H SCALE SUPPORT, AFFINE VIEWS, AND OBSERVATION DIVERSITY

We fix 32 previously exposed cohort-B families at indices $0, 8, \ldots, 248$, three noise levels, and four evaluation views: identity, the original $[-6, 6]$ log-scale draw, a wider draw obtained by multiplying its log factors by 1.5, and the original scale draw with affine offsets. Each offset is a fixed uniform fraction in $[-1, 1]$ of the corresponding visible input or output RMS. The fraction is shared across

Table 12: **Realized numerical work under a shared policy.** Expression forwards per complete return on the 32-family follow-up. Each row has 576 queries. The equal proposal count and fitting policy permit different realized work across output types.

Target	Request	Median forwards	Mean forwards	90th percentile
Literal	F	99,904.0	136,264.5	297,283
Literal	C	128,761.0	172,316.5	375,971
Roles	F	105,140.0	162,323.2	397,538
Roles	C	144,081.5	180,319.4	398,332

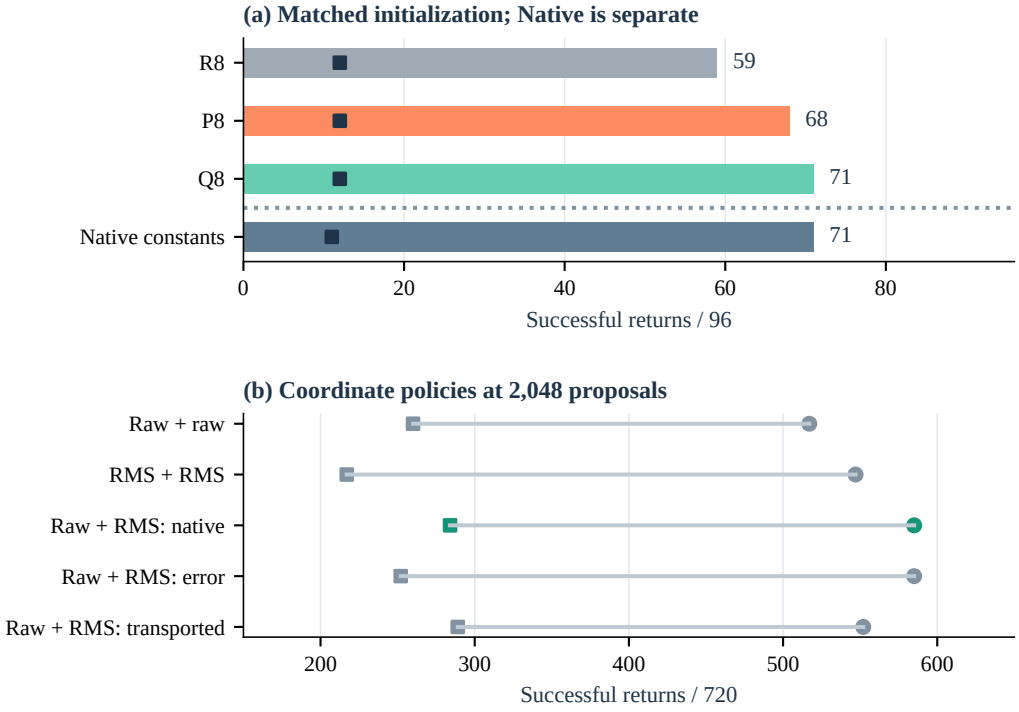


Figure 15: **Numerical information improves the returns of two controlled searches.** (a) Same parents and eight start positions: random (R8), one predicted start (P8), or its coarse signed magnitudes (Q8). The dotted line separates the native constants reference with its historical fitting policy. (b) Frozen120 searches on 720 catalogue conditions, with 2,048 proposals per policy. In displayed order, the five policies retain 5, 5, 5, 6, and 3 nonfinite returns; every complete mean is infinite. Bars and circles show high-accuracy counts; squares show strict successes. In (b), segments connect the two endpoint counts for each policy. The full coordinate table reports complexity and risk for all five policies.

noise strata. Every query has 128 visible points and 64 proposals; all arms use native NumPy refinement and common final formula evaluators.

The training comparison retains the original 4,096 families, observation sets, family order, 2,000 updates, 16 view presentations per update, and final checkpoint policy. Three training seeds, two capacities, and two target formats are used throughout. New variants narrow the original nonidentity log factors to $[-3, 3]$, widen them to $[-9, 9]$, or add affine offsets while retaining the identity view. A separate RMS-canonical variant presents the same RMS-normalized observation set in all four views and uses RMS inference with formula transport. It is compared with scale training followed by the same RMS inference, alongside raw-coordinate inference. All variants retain the native target-length

limit; when a transformed target exceeds that limit or fails numerical transport, both supervision formats retain the unshifted scale view and record the fallback.

What the canonical control makes invariant. Let $r_j = \sqrt{\frac{1}{n}\sum_i x_{ij}^2}$ and $r_y = \sqrt{\frac{1}{n}\sum_i y_i^2}$ be nonzero visible RMS values. For positive rescalings $x'_j = a_j x_j$, $y' = by$, the normalized observations are identical in exact arithmetic because $r'_j = a_j r_j$ and $r'_y = b r_y$. With the same random stream, a fixed search and selector acting on these canonical observations therefore return the same canonical formula g; transport gives

$$\hat{f}'(x') = b r_y \, g\left(\frac{x'_1}{a_1 r_1}, \ldots, \frac{x'_d}{a_d r_d}\right) = b\hat{f}(x'/a).$$

This control tests positive scale equivariance of the complete normalized route. Floating-point trajectories and scores recomputed after transport depend on the execution coordinates. Zero-RMS columns use the recorded unit-scale convention.

A non-scale augmentation control. To distinguish coordinate diversity from additional observation diversity, we train a further control on the same families and update budget. Its first view keeps the original observations; the remaining three independently resample points and noise at the original coordinates. Targets remain the identity targets. Thus each family receives four observation views without new coordinate scales. The scale/range/RMS comparison contains 64,512 returns and this additional control contains 9,216. All conditions are retained.

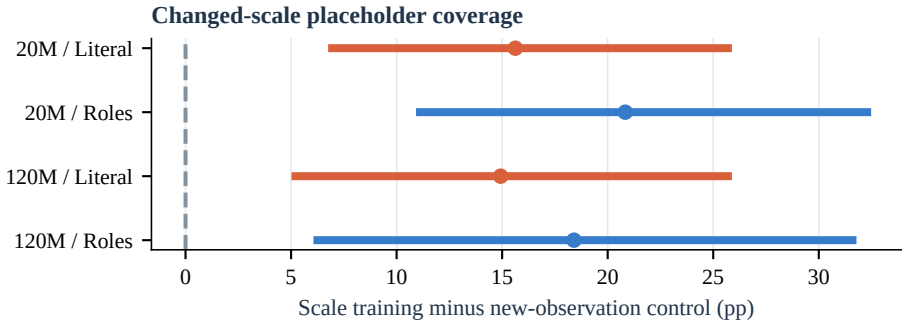


Figure 16: **Scale diversity adds coverage beyond resampled observations.** Placeholder output on the changed-scale view, 32 fixed families and three training seeds. Points show scale training minus independent original-scale observation resampling; bars are pointwise 95% family–seed bootstrap intervals. This is a secondary analysis on exposed families.

Scale training exceeds observation resampling by 15.63 and 20.83 percentage points at 20M (literal and role targets), and 14.93 and 18.40 at 120M. The corresponding intervals are $[6.94, 25.69]$, $[11.11, 32.29]$, $[5.21, 25.69]$, and $[6.25, 31.60]$. Resampling alone differs from identity continuation by $+0.69, +0.69, +0.35, -2.08$ points in these four cells; each interval includes zero. These controls support a role for coordinate diversity beyond the added variation in observed points and noise.

Figures 17 and 18 report every training policy, inference coordinate, evaluation view, target format, and capacity. Each cell pools 288 returns (32 families, three noises, three training seeds). The diagrams retain the original and changed coordinates separately; the paired machine-readable summaries additionally provide the $R^2 \geq .999$ endpoint, complete risk, nonfinite counts, complexity, and expression calls. The view-specific comparisons are exploratory, with pointwise intervals. The prespecified cohort-B interaction has its own adjusted interval.

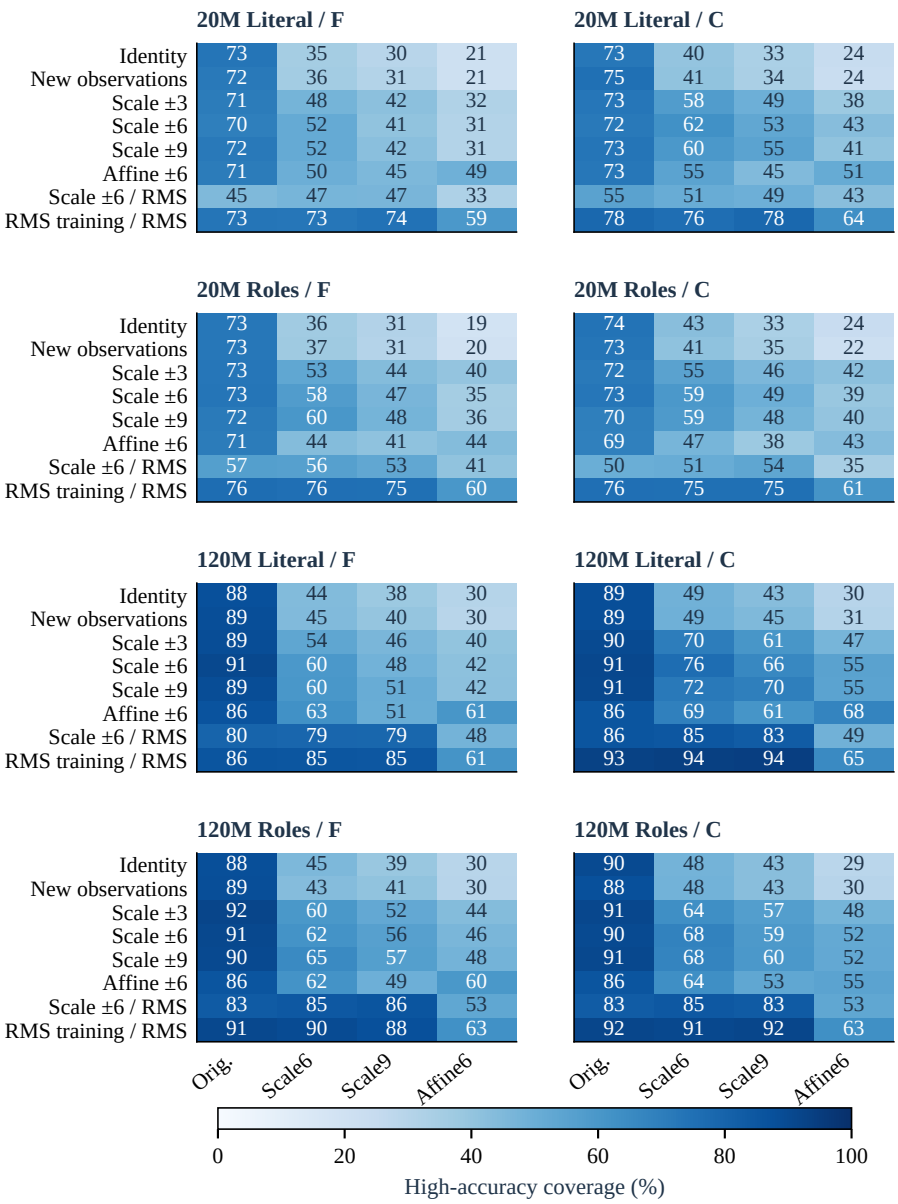


Figure 17: **Coverage across coordinate interventions.** All values are percentages at $R^2 \geq .99$. Rows give training and inference policies; columns within each panel give evaluation coordinates. "New observations" is the non-scale resampling control. RMS training uses the same canonical view four times. The two panel columns request placeholders/constants; the four panel rows identify capacity and training target.

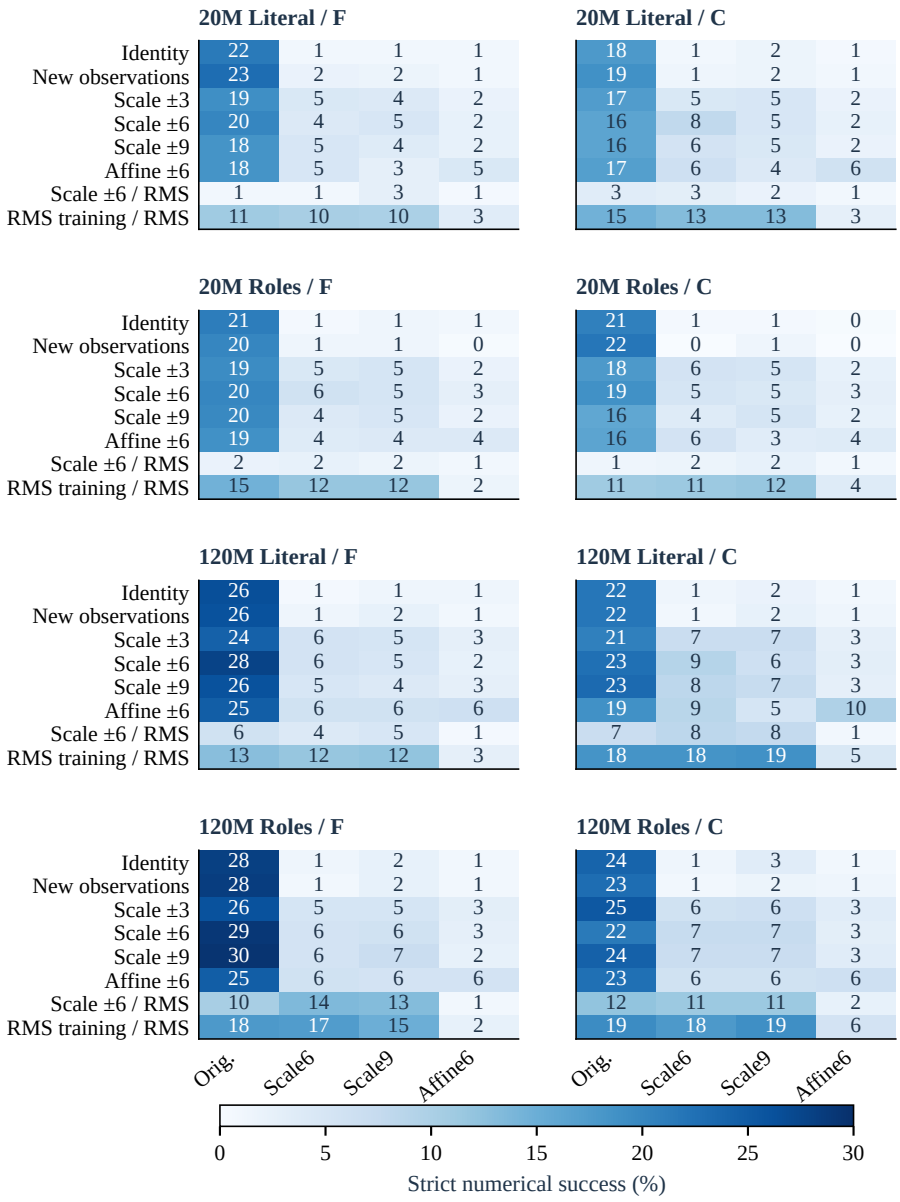


Figure 18: **Strict numerical success across the same interventions.** The design matches Figure 17; the endpoint is $R \leq 2^{-23}$. Every percentage retains the complete cell denominator. Its separately labeled 0–30% color scale makes the strict range visible.

Adjacent neural-surrogate approaches. Separability-based symbolic search (Ibata et al., 2026) and network-dynamics discovery with physical priors (Hu et al., 2025) illustrate other uses of learned numerical representations in equation search. These complement the amortized generation methods surveyed by Palakonda et al. (2026); our experiments address target format and the subsequent numerical interface within the specified generators.

I EXTERNAL BASELINE CONFIGURATIONS

The public comparison uses 24 datasets and three official splits per system. All 72 runs of each baseline use the settings below. Each system requests four CPU threads and a native 180-second search target. The budget applies to internal search. Observed total service ranges are 180.35–184.48 seconds for PySR, 180.01–180.36 for Operon, and 179.04–182.84 for TIR.

Table 13: Public baseline search, complexity, and return-selection settings. Standardization is fitted on the training split and uses population standard deviations; held-out targets never enter preprocessing or selection.

System	Coordinates	Search and complexity limits	Selection
PySR	Center and standardize every input and the target	31 populations of 27; at most 100,000 iterations; maximum expression size 30; 64-bit precision	Native `best` rule
Operon	Center and standardize every input and the target	Population 1,000; at most 10^6 generations and 10^9 evaluations; length 30, depth 10; 10 optimizer iterations	Objectives: R^2 and length; Bayesian information criterion for the return
TIR	Original raw coordinates	Population 1,000; at most 100,000 generations; crossover/mutation 0.3/0.7; exponents in $[-5, 5]$; no separate tree-size cap recorded	Native multiobjective return; R^2 objective; zero penalty and zero extra numerical iterations

Table 14: Configured expression vocabulary. `Id` is the identity transform. Operator sets and structural restrictions remain system-specific.

System	Operators or input transforms	Additional restrictions or target transforms
PySR	$+, -, \times, /, \text{power}$; $\sin, \cos, \exp, \log, \sqrt{\cdot}, \lvert\cdot\rvert$	Power exponent child has size at most one; no additional base-child size constraint
Operon	$+, -, \times, /$; constants and variables; $\sin, \cos, \exp, \log, \sqrt{\cdot}, \text{square}$	Length and depth limits from Table 13
TIR	`Id`, $\tanh, \sin, \cos, \log, \exp, \sqrt{\cdot}$	Target transforms: `Id`, $\arctan, \tan, \tanh, \log, \exp, \sqrt{\cdot}$

The separate 600-second PySR/Operon controls use the corresponding native recipes on 12 fixed synthetic conditions, with all 128 visible observations and four threads. Appendix J.3 reports these runs with their synthetic task set and budget.

J ADDITIONAL STRUCTURAL AND NUMERICAL CONTROLS

J.1 GENERALIZATION ACROSS SERIALIZED STRUCTURAL MOTIFS

The 48-family panel distinguishes three mutually exclusive strata of 16: all local operator windows and variable-coupling motifs seen in continuation training; at least one unseen operator window;

and seen windows with an unseen coupling motif. Membership uses the realized target in its evaluated coordinates and all 16,384 serialized continuation views. Each family has two independently sampled sets of 128 visible points and three noise levels. Within each stratum, eight families use identity coordinates and eight use a changed coordinate system; the two samples of a family share its scale. Three training seeds and four target–output configurations give 3,456 returns, or 864 per pooled cell.

Roles give 601 versus 584 high-accuracy placeholder returns; literals give 632 versus 608 with constant emission. The high99 interaction is -4.75 points, with pointwise 95% interval $[-12.85, 2.55]$ and nine-comparison interval $[-16.67, 5.56]$. Two seed interactions are negative and the third is zero. At $R^2 \geq .999$, the interaction is -7.99 points: its pointwise interval $[-15.63, -1.27]$ excludes zero, while its adjusted interval $[-19.44, 1.04]$ crosses zero. The strict interaction is -2.55 points, with both intervals crossing zero. This structural panel therefore gives the same pooled high99 direction with less resolved uncertainty than the 256-family confirmation.

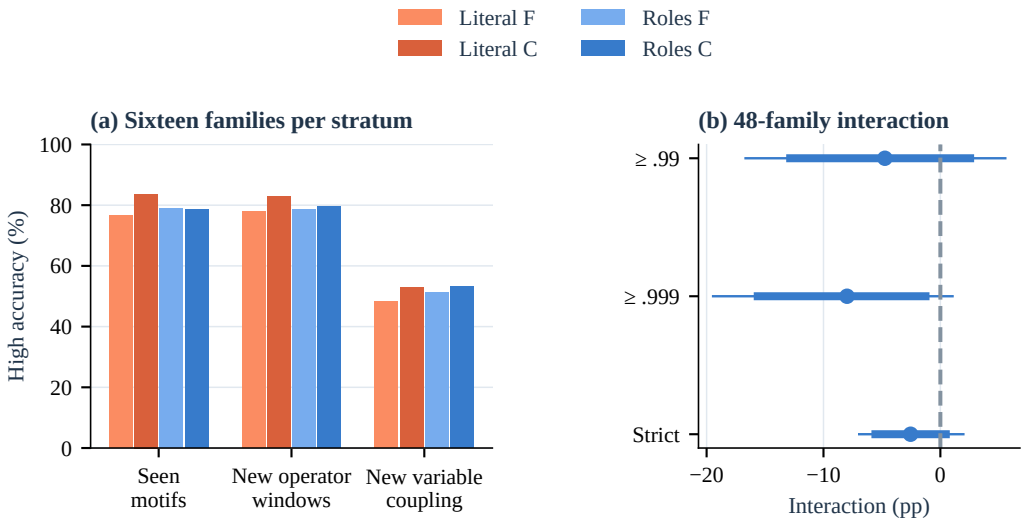


Figure 19: Structural-motif controls. (a) Each bar pools 288 conditions from 16 families, two observation samples, three noise levels, and three seeds. (b) The pooled 48-family interactions retain crossed family–seed uncertainty; light intervals adjust the nine endpoint/contrast comparisons. F/C denote placeholder/constant emission.

Table 15: All 48-family returns, 864 conditions per row. Strict uses FVU $\leq 2^{-23}$. NF remains in the denominator and complete mean; there are no empty returns. The single nonfinite role/constant case belongs to the seen-motif stratum.

Target / output	$R^2 \geq .99$	$R^2 \geq .999$	Strict	NF	Mean FVU
Literal / F	584	251	38	0	0.03290
Literal / C	632	323	64	0	0.02850
Roles / F	601	289	44	0	0.02608
Roles / C	608	292	48	1	∞

J.2 Predicted, perturbed, and cross-architecture initial values

The Flash-ANSR controls retain the original constant-emission parent pools on 96 old development conditions from 16 families. Every parameterized fit has eight initialization positions: R8 uses eight shared Gaussian starts; P8 or Q8 replaces one with predicted coefficients or their signed

nearest powers of two. Z8 substitutes zeros; S8 substitutes another expression's whole coefficient vector at matched free-parameter count; M8 substitutes independent random signs times 10^U, $U \sim \mathcal{U}[-0.5, 3]$, from the native magnitude-span rule. S8 changes the expression-family donor for 4,856 of 4,931 numerical parents; 75 unavoidable identity donors remain. The seven other starts, parameter roles, 200-forward/five-second per-fit limits, and common refinement remain fixed.

The official E2E model supplies a second initialization setting. Its 64 decoded expressions yield up to eight stable-unique parents in draw order on the same 96 tasks. R8/P8/Q8 fit each retained parent with the native E2E expression, analytic-gradient BFGS, and the same 200-forward/five-second limits per start; selection uses visible MSE. This is an initialization intervention in another model family, with a different parent pool and solver from Flash-ANSR.

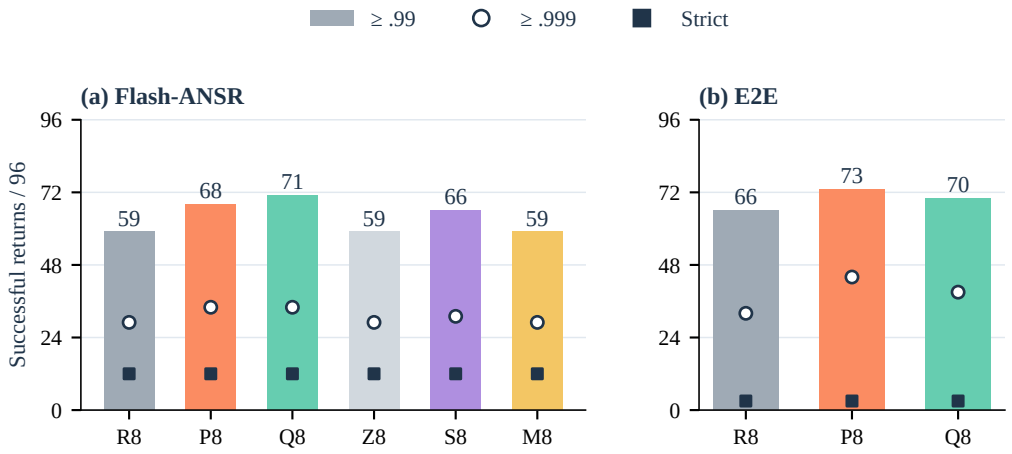


Figure 20: Numerical initialization on 96 fixed conditions. Bars count high99 returns, hollow circles high999 returns, and black squares strict numerical successes. Comparisons hold the parent pool fixed within each system. R/P/Q denote random, predicted, and coarse predicted starts; Z/S/M denote zero, shuffled, and native magnitude-span replacements.

Table 16: Complete initialization outcomes, 96 conditions per row. All returns are finite and nonempty; means retain every condition.

System	Start policy	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Mean FVU
Flash-ANSR	R8	59	29	12	0.05075
Flash-ANSR	P8	68	34	12	0.03889
Flash-ANSR	Q8	71	34	12	0.03795
Flash-ANSR	Z8	59	29	12	0.04877
Flash-ANSR	S8	66	31	12	0.04155
Flash-ANSR	M8	59	29	12	0.05000
E2E	R8	66	32	3	0.98660
E2E	P8	73	44	3	0.01898
E2E	Q8	70	39	3	0.04870

E2E predicted initialization raises high99 coverage from 66 to 73, or $+7.29$ points (pointwise $[2.08, 13.54]$; nine-comparison $[0.00, 16.67]$). Its high999 gain is $+12.50$ points, with adjusted interval $[2.08, 25.00]$. Coarse E2E starts reach 70 high99 returns, and all three have three strict successes. In Flash-ANSR, shuffling gives 66 high99 returns versus P8's 68; zero and native magnitude-span starts each give 59. The S8-minus-R8 gain is $+7.29$ points (pointwise $[2.08, 14.58]$; 18-comparison $[0.00, 18.75]$). The zero/magnitude-span difference from P8 is -9.38 points (point-

wise $[-17.71, -2.08]$; adjusted $[-21.88, 0.00]$). All Flash policies retain 12 strict successes. Predicted values improve both fixed-parent pipelines, with coarse and shuffled values retaining much of the Flash-ANSR high99 gain.

J.3 Proposal and expression-call budgets

The compute panel retains two fixed conditions for each of 16 exposed families. Each target–output setting independently decodes 16, 64, or 256 proposals, with total expression-call caps 1,000, 4,000, or 16,000 and soft refinement limits 15, 60, or 240 seconds. Calls include numerical-Jacobian probes, failed calls, post-fit evaluation, and refinement fits. A cutoff retains completed candidates; if none remains, the visible-mean fallback stays in the denominator. Each search selects its winner from the candidates completed within its own budget.

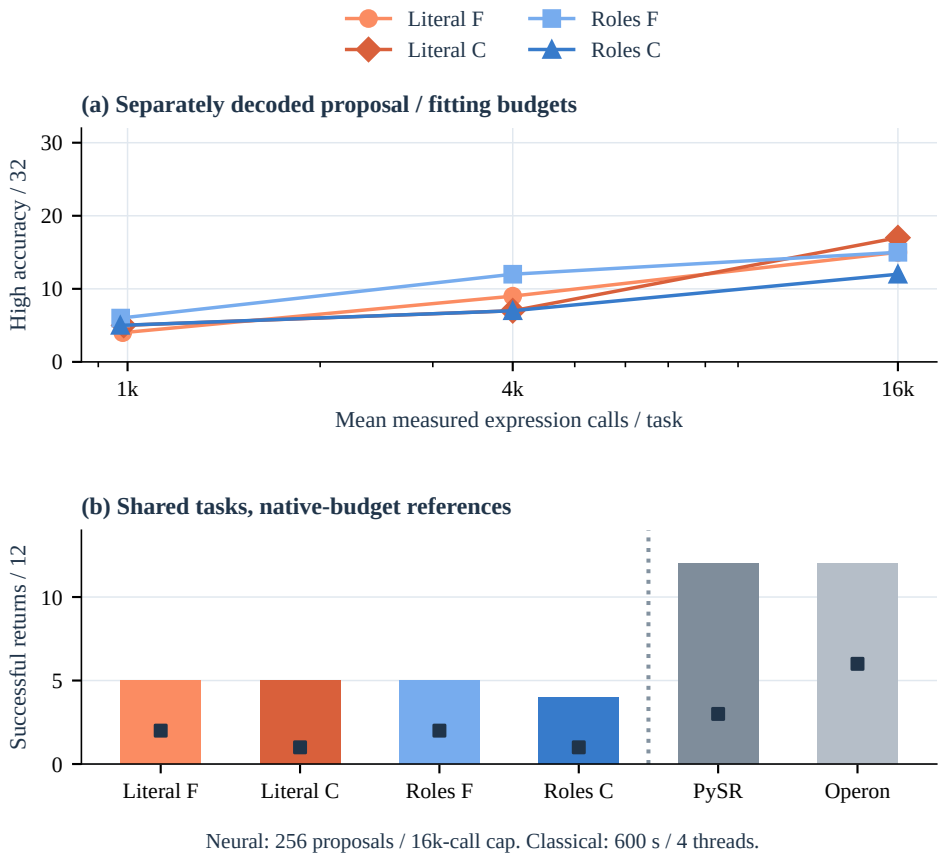


Figure 21: Budgeted search and native classical references. (a) All 32 neural conditions at three independently decoded proposal budgets, plotted against measured expression calls. (b) The same 12 synthetic conditions and visible observations are used for every shown system. Neural endpoints use 256 proposals and a 16,000-call cap; PySR/Operon use their native 600-second, four-thread budgets. Bars count high99 and squares strict success. The two budget units remain distinct.

High99 counts rise with budget in all four neural settings. The nine-comparison high99 interaction intervals include zero at every budget. On the fixed 12 synthetic conditions, PySR/Operon reach 12/12 high99 and 3/6 strict successes, with complete mean FVU 0.000688/0.000181. Neural 256-proposal endpoints reach 4–5 high99 and 1–2 strict successes, with means 0.15884–0.33894; all these 12-task returns are finite and nonempty. Classical median services are 600.47/600.20 seconds. Neural records provide generation and refinement/build phases but no complete service-time field, so the neural compute axis uses expression calls. This panel is distinct from the 72-task public comparison.

Table 17: All 32-condition neural budget cells. Mean calls count actual expression evaluations per condition. Empty selections use the visible mean; all reported predictions are finite.

Target / output	Draws	Mean calls	$R^2 \geq .99$	Strict	Empty	Mean FVU
Literal / F	16	982.8	4	1	14	0.62923
Literal / F	64	4000.0	9	5	3	0.26130
Literal / F	256	16000.0	15	7	0	0.14766
Literal / C	16	986.3	5	2	13	0.50167
Literal / C	64	4000.0	7	3	2	0.25664
Literal / C	256	16000.0	17	6	0	0.11456
Roles / F	16	975.5	6	2	14	0.49728
Roles / F	64	4000.0	12	4	2	0.21817
Roles / F	256	16000.0	15	7	1	0.17342
Roles / C	16	974.2	5	2	15	0.58450
Roles / C	64	4000.0	7	3	6	0.89310
Roles / C	256	16000.0	12	5	0	0.19990

J.4 Autoregressive responses to emitted values

Changing FREE coefficient bytes during decoding changes subsequent operator and variable sequences. This panel uses the literal/scale 20M checkpoint on 32 exposed conditions: 16 families, two noise levels, and the changed-coordinate view. Each condition has 32 draws. After a completed FREE value span, the intervention replaces its eight value bytes with either a signed nearest power of two or the next same-role value from a fixed cyclic donor stream. Delimiters, typed numbers, and already emitted structural tokens are unchanged. All three modes then use the same cold eight-start fit, with predicted initialization removed and the native typed rules, refinement, and selection retained.

Table 18: Raw decoding and final returns. Raw validity is measured before simplification and retains all 1,024 draws per mode. Changed O/V counts operator or variable sequence changes against the original draw index, after reversing the manual byte edits. The three accuracy columns and complete mean use all 32 final conditions per mode; all returns are finite and nonempty.

Mode	Valid / invalid	Changed O/V	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Mean FVU
Original	917 / 107	—	12	7	1	0.09168
Coarse	950 / 74	841	13	5	1	0.09463
Cyclic donor	924 / 100	743	16	4	1	0.09152

Coarse and cyclic interventions change 3,825 and 4,605 value spans. Subsequent operator or variable tokens differ in 841 and 743 of the 1,024 paired draws; 47 and 81 pairs differ solely in the manually edited values. All raw terminals, including invalid sequences, enter these counts. Changed termination lengths alter the active batch and subsequent random-number assignment, even with a shared batch seed. The counts measure the full decoding response to the intervention. Final high99 counts rise from 12 to 13/16, while high999 counts fall from 7 to 5/4 and strict successes remain one.

J.5 Constant emission on the exposed physical panel

For 120M models, scale training increases high99 physical returns in all four target–output settings (Table 19). The panel contains 12 exposed physical cases, two fixed noisy observations per case, and three training seeds, giving 72 returns per cell. Every search uses 1,024 proposals in raw coordinates and the native fitting and selection path. The comparison includes the neural return itself, without guarded or library candidate combinations.

Constant-minus-placeholder high99 changes are 0.00, -1.39, $+5.56$, and -1.39 points for literal/identity, literal/scale, roles/identity, and roles/scale, respectively. Their pointwise 95% intervals are $[-5.56, 6.94]$, $[-8.33, 0.00]$, $[-5.56, 20.83]$, and $[-8.33, 0.00]$: all include zero. The 10,000 crossed case–seed resamples keep both observations of each case together. Role/identity constants

Table 19: All 576 physical returns. F/C denote placeholder/constant emission. Each row contains the same 72 case–observation–seed combinations. NF and empty counts are zero in every row; complete mean FVU retains large finite errors. Median bits measure returned expression length.

Target	Training	Output	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Mean FVU	Median bits
Literal	Identity	F	58	48	23	0.73985	45.46
Literal	Identity	C	58	47	21	0.61045	45.46
Literal	Scale	F	62	58	25	5.69310	35.86
Literal	Scale	C	61	53	24	3.96259	38.10
Roles	Identity	F	53	49	23	4.54×10^{10}	45.23
Roles	Identity	C	57	49	20	1.19×10^{14}	46.04
Roles	Scale	F	61	58	28	41.82038	33.25
Roles	Scale	C	60	57	28	10.05483	30.64

add six high99 returns and lose two, but reduce strict successes from 23 to 20. Their mean FVU is 1.19×10^{14}, driven by a finite error of 8.54×10^{15}. Coverage, strict numerical success, and tail risk thus describe different changes in the returned expressions.

K Alternative continuous solver and refinement removal

The four-pool mechanism panel in Appendix AE.1 is repeated with the native trust-region reflective least-squares optimizer (TRF), replacing Levenberg–Marquardt (LM) in both continuous fitting and internal refinement fits. The same 32 exposed conditions from six families, first eight parent positions per source pool, parameter identities, and R4/P4 start vectors are retained. Both solvers have a 200-expression-forward and five-second cap per fit, including numerical-Jacobian probes. The canonical continuous parent and the final refinement result are separately scored by the same visible native rule. The refinement-removal control selects canonical continuous parents using the same visible score.

Table 20: **Alternative solver and discrete refinement on fixed pools.** Every count uses all 32 conditions. The last four columns use refined final returns. Both solvers retain the same four starts per parameterized fit and the same caps; all selected predictions here are finite and nonempty.

Solver	Pool / starts	Before refine. high99	After refine. high99	High999	Strict	Mean FVU
LM	literal / F / R4	11	11	1	1	0.0921
LM	literal / C / R4	10	11	2	1	0.1627
LM	roles / F / R4	8	9	4	3	0.1020
LM	roles / C / R4	6	9	2	1	0.2063
LM	literal / C / P4	10	12	2	1	0.0980
LM	roles / C / P4	6	9	2	1	0.2235
TRF	literal / F / R4	11	11	1	1	0.1196
TRF	literal / C / R4	10	11	2	1	0.2082
TRF	roles / F / R4	8	9	4	3	0.1321
TRF	roles / C / R4	6	9	2	1	0.2364
TRF	literal / C / P4	10	11	2	1	0.1409
TRF	roles / C / P4	6	9	2	1	0.1672

Under shared random-start fitting, both optimizers give 11/11/9/9 high-accuracy final returns for literal F/C and role F/C. Their cold interface-interaction point estimate is zero. Predicted-start replacement raises literal/C from 11 to 12 under LM and leaves it at 11 under TRF; the role/C count remains nine. Before discrete refinement, both cold interactions are -3.13 points. Six-family intervals remain wide: TRF's final cold and predicted-start interactions both have pointwise interval $[-19.44, 18.75]$. The interval supports both directions of interaction on this restricted candidate pool.

All 1,024 selected parent positions remain in the record. The TRF execution uses 3,032 distinct continuous fits after exact shared-input reuse: 2,281 return natively, 643 stop at a budget with the

best observed finite state retained, and 108 fail. All 1,245 distinct postprocessing calls complete. The six policies are charged their full logical fitting work despite physical reuse. This single-training-seed comparison uses unequal numbers of conditions across six exposed families and eight parents per pool. The main cohort-B comparison uses all 64 proposals.

L THE CROSSOVER ACROSS A CONTINUOUS PRECISION PROFILE

Figure 22 re-evaluates the same 27,648 frozen cohort-B returns over tolerances from 10^{-1} to 10^{-8}. Each coverage curve has the full 4,608-condition denominator, including nonfinite and empty returns. The interaction uses the original matched family–training-seed bootstrap indices. This descriptive sweep supplements the three prespecified endpoints. Shading shows pointwise intervals; the endpoint markers retain their multiplicity-adjusted intervals.

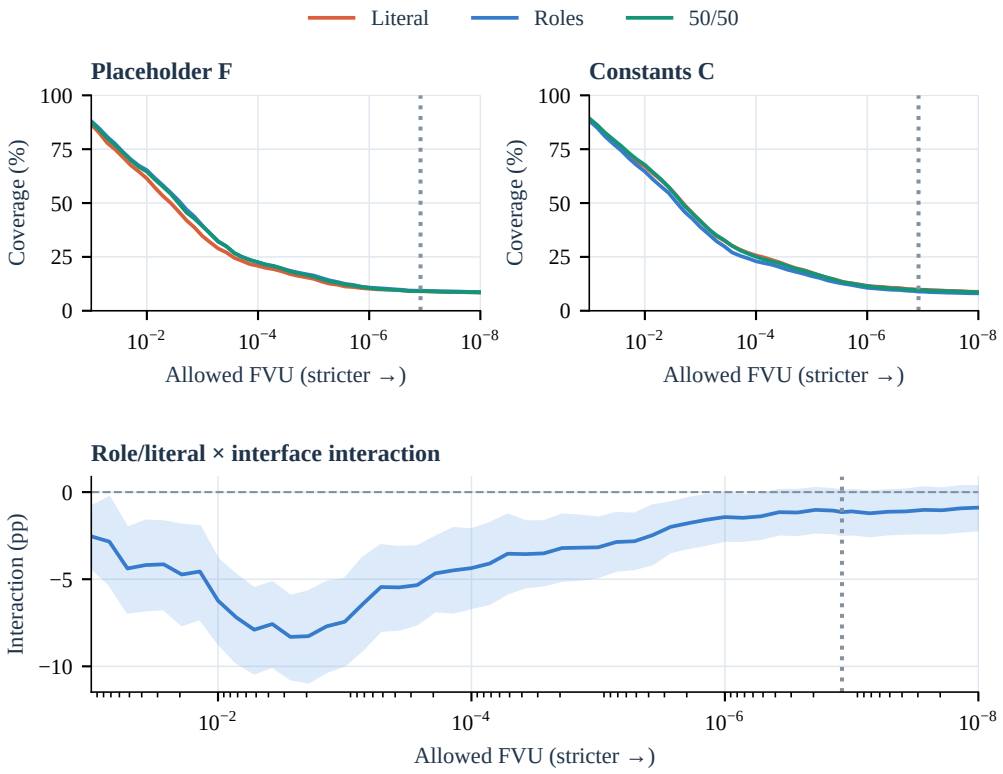


Figure 22: **The interface interaction across predictive precision.** Top panels: coverage of the fixed formulas under progressively stricter FVU cutoffs. Bottom: role-minus-literal effect under constants minus that under placeholders; shading is the pointwise 95% paired bootstrap interval. The vertical dotted line marks the strict 2^{-23} endpoint.

At $R^2 \geq .9999$, literal F/C has 960/1,183 successes and roles F/C has 1,036/1,058, giving an interaction of -4.36 points with pointwise interval $[-6.64, -2.15]$. At $R^2 \geq .99999$, the respective counts are 683/816 and 754/741: interaction -3.17 points, $[-4.86, -1.48]$. Thus the observed crossover extends across several orders of predictive error before its uncertainty includes zero near the strict numerical endpoint. Balanced F/C counts at the two intermediate thresholds are 1,037/1,155 and 730/805. The original high99, high999, and strict effects and intervals are reproduced exactly by this saved-return calculation.

M BALANCED SUPERVISION AT 120M ON COHORT A

The fixed 50/50 literal–role-masked recipe was trained for 2,000 updates at 120M with the same three training seeds as the specialized controls. All 9,216 new returns use 64 proposals per task and the unchanged native refinement and selection rules. This follow-up reuses the previously exposed 256-family cohort A; each recipe–interface cell contains 4,608 conditions (three seeds, two observation views, and three noise levels).

Table 21 reports all three accuracy thresholds, nonfinite conditions, and the full-denominator median risk. Balanced supervision gives 3,493 placeholder and 3,778 constants returns with $R^2 \geq .99$, compared with 3,484 for the role-masked placeholder control and 3,755 for the literal constants control. The paired differences are $+0.20$ and $+0.50$ percentage points, with nominal 95% intervals $[-1.17, 1.43]$ and $[-0.63, 1.61]$; both adjusted intervals also include zero. Other endpoints retain mixed point directions: balanced placeholder has 14 fewer $R^2 \geq .999$ successes than the role-masked control, and balanced constants has 17 fewer strict successes than the literal control.

Table 21: 120M on cohort A. Each row has $n = 4,608$; F/C denote placeholder/constants output. Counts are $R^2 \geq .99$, $R^2 \geq .999$, and strict FVU $\leq 2^{-23}$. NF counts conditions with at least one nonfinite prediction. Medians retain the full denominator. All native returns are nonempty. Full mean FVU includes every return and is infinite in every row except literal F (0.02898).

Supervision	Interface	$\geq .99$	$\geq .999$	Strict	NF	Median FVU
Literal	F	3389	2239	852	0	1.11×10^{-3}
Literal	C	3755	2694	911	1	5.13×10^{-4}
Roles	F	3484	2419	873	1	8.46×10^{-4}
Roles	C	3579	2427	795	1	8.19×10^{-4}
Balanced	F	3493	2405	894	2	8.72×10^{-4}
Balanced	C	3778	2713	894	1	4.96×10^{-4}

Table 22 retains all 12 registered endpoint comparisons. Resampling crosses the three training seeds with the same 256-family resample, retaining all six within-family conditions (10,000 replicates). After the 12-comparison adjustment, the intervals remain positive for balanced versus literal F at $R^2 \geq .999$ and for balanced versus role-masked C at all three thresholds. All comparisons with the stronger specialist in each interface include zero.

Table 22: Paired balanced-minus-specialist differences in percentage points on cohort A. Nominal 95% and Bonferroni-12 adjusted percentile intervals use the same crossed seed/family resamples.

Comparator	Endpoint	Δ (pp)	Nominal 95%	Adjusted interval
Literal F	$R^2 \geq .99$	$+2.26$	$[0.46, 4.14]$	$[-0.33, 5.08]$
Literal F	$R^2 \geq .999$	$+3.60$	$[1.65, 5.49]$	$[0.65, 6.38]$
Literal F	Strict	$+0.91$	$[-0.39, 2.15]$	$[-1.15, 2.80]$
Roles F	$R^2 \geq .99$	$+0.20$	$[-1.17, 1.43]$	$[-2.02, 2.15]$
Roles F	$R^2 \geq .999$	-0.30	$[-1.91, 1.32]$	$[-2.73, 2.21]$
Roles F	Strict	$+0.46$	$[-1.06, 2.04]$	$[-2.02, 3.11]$
Literal C	$R^2 \geq .99$	$+0.50$	$[-0.63, 1.61]$	$[-1.24, 2.21]$
Literal C	$R^2 \geq .999$	$+0.41$	$[-1.09, 1.89]$	$[-1.91, 2.59]$
Literal C	Strict	-0.37	$[-1.82, 0.98]$	$[-2.55, 1.79]$
Roles C	$R^2 \geq .99$	$+4.32$	$[2.65, 6.14]$	$[1.89, 7.04]$
Roles C	$R^2 \geq .999$	$+6.21$	$[3.78, 8.66]$	$[2.59, 9.83]$
Roles C	Strict	$+2.15$	$[0.76, 3.58]$	$[0.07, 4.23]$

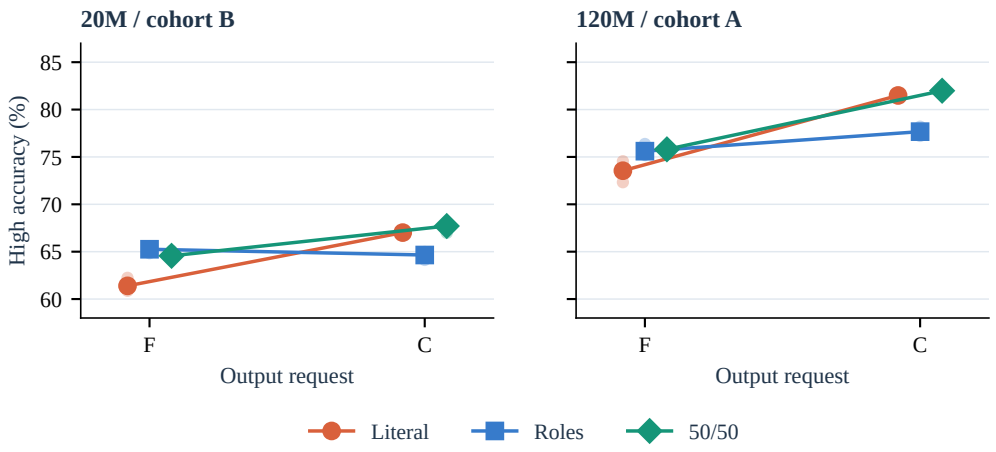
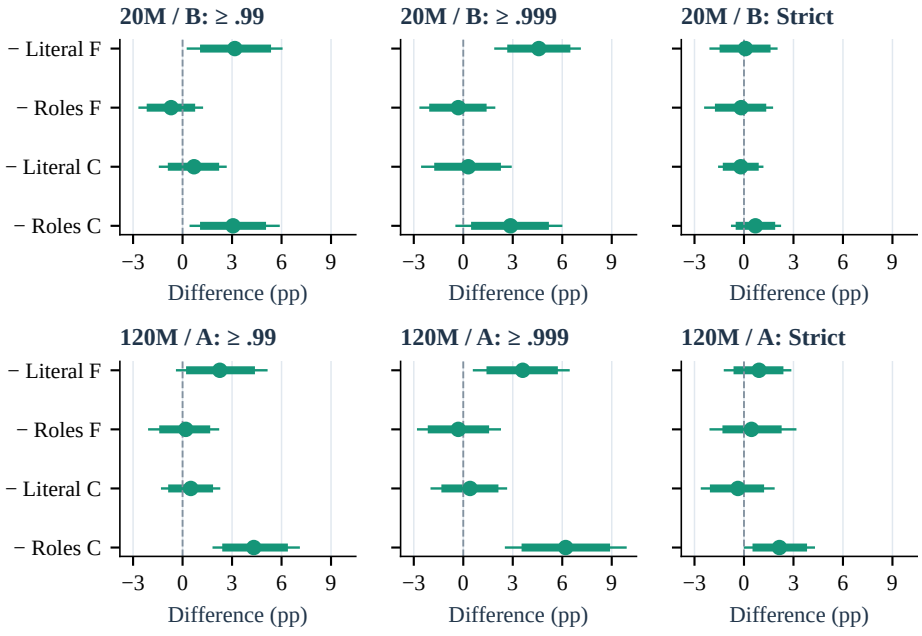


Figure 23: The supervision–interface profiles are shown separately for 20M on cohort B and 120M on cohort A. Large symbols join pooled $R^2 \geq .99$ rates ($n = 4,608$ per cell); small symbols are the three individual training seeds, not confidence intervals. Both panels use 64 proposals per return. The two panels describe different combinations of model capacity and family cohort. Paired 120M uncertainty is given in Table 22.



Balanced minus each specialist. Thick: nominal 95%; thin: adjusted 12.

Figure 24: **Balanced training across interfaces, precision, and capacity.** Each row compares balanced training with the named specialist under the same output request. The top row uses the original three-seed cohort-B comparison; the bottom row is the 120M follow-up on exposed cohort A. Thick intervals are nominal 95%; thin intervals use each study's twelve-comparison adjustment. All counts and risk summaries remain in the corresponding tables.

N NATIVE E2E SUPERVISION AND INITIALIZER FOLLOW-UP

We continued the official 93.45M-parameter E2E checkpoint for 2,000 updates with literal or native skeleton targets and three paired training seeds. Each checkpoint received the same 16,000 observation views (128 points per view), using E2E's native token-mean loss. The six last-update checkpoints were evaluated on 16 exposed development families with six conditions each. Each rollout retained all 64 draws and fitted the first eight valid unique parents. Both refinement routes requested eight starting slots: R_8 used Gaussian starts, and P_8 replaced one start with the emitted numeric vector when available. Each fit retained its predeclared 200-forward/5-second budget and best finite state on budget termination.

Table 23 reports all 1,152 returns. Numeric initializers were available for 2,246/2,304 literal-trained parents and 0/2,304 skeleton-trained parents; consequently, the skeleton P_8 route is exactly the prescribed R_8 fallback. This experiment compares training targets and initialization within E2E's single native generation interface.

Table 23: Native E2E follow-up on the exposed development panel. Each row has $n = 288$. Strict means FVU $\leq 2^{-23}$. All final predictions are finite and all returns are nonempty.

Training target	Refinement	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Median FVU
Literal	R_8	178	89	6	2.98×10^{-3}
Literal	P_8	189	126	12	1.59×10^{-3}
Native skeleton	R_8	160	83	4	3.78×10^{-3}
Native skeleton	$P_8 = R_8$ fallback	160	83	4	3.78×10^{-3}

The literal initializer gain at the primary $R^2 \geq .99$ threshold is $+3.82$ percentage points (13 added, 2 lost); its nominal interval includes zero. The finer $R^2 \geq .999$ endpoint increases by $+12.85$ points (40 added, 3 lost), with a positive Bonferroni-15 interval. Strict recovery increases from 6 to 12, with an interval touching zero. Table 24 retains all five contrasts at all three thresholds. Its interaction equals the literal initializer effect because skeleton P_8 and R_8 execute the same starts.

Table 24: Differences in percentage points. The same 10,000 crossed resamples of three training seeds and 16 families retain all six within-family conditions. Adjusted intervals use Bonferroni-15 over the five contrasts and three endpoints. L/S denote literal/native-skeleton training; $I = (L_{P_8} - L_{R_8}) - (S_{P_8} - S_{R_8})$.

Contrast	Endpoint	Δ (pp)	Nominal 95%	Adjusted interval
L: $P_8 - R_8$	$R^2 \geq .99$	$+3.82$	$[-0.35, 8.68]$	$[-4.17, 11.81]$
	$R^2 \geq .999$	$+12.85$	$[6.94, 19.44]$	$[4.17, 23.26]$
	Strict	$+2.08$	$[0.00, 5.56]$	$[0.00, 7.64]$
S: $P_8 - R_8$	$R^2 \geq .99$	0.00	$[0.00, 0.00]$	$[0.00, 0.00]$
	$R^2 \geq .999$	0.00	$[0.00, 0.00]$	$[0.00, 0.00]$
	Strict	0.00	$[0.00, 0.00]$	$[0.00, 0.00]$
L–S at R_8	$R^2 \geq .99$	$+6.25$	$[-0.69, 13.89]$	$[-4.17, 19.79]$
	$R^2 \geq .999$	$+2.08$	$[-6.60, 11.46]$	$[-11.81, 18.87]$
	Strict	$+0.69$	$[-1.39, 4.17]$	$[-2.89, 6.25]$
L–S at P_8	$R^2 \geq .99$	$+10.07$	$[2.78, 18.75]$	$[0.00, 25.35]$
	$R^2 \geq .999$	$+14.93$	$[5.90, 25.35]$	$[1.27, 32.29]$
	Strict	$+2.78$	$[0.00, 6.25]$	$[0.00, 9.38]$
I	$R^2 \geq .99$	$+3.82$	$[-0.35, 8.68]$	$[-4.17, 11.81]$
	$R^2 \geq .999$	$+12.85$	$[6.94, 19.44]$	$[4.17, 23.26]$
	Strict	$+2.08$	$[0.00, 5.56]$	$[0.00, 7.64]$

Native and independent prediction checks agree for every return. The 39,110 physical fits comprise 22,420 converged starts, 14,657 forward-cap stops, 1,842 other optimizer stops, and 191 invalid-variable attempts. The latter reference variables outside the query's input dimension. All final predictions are finite and nonempty. Fitting incurs 4,690,689 expression forwards; postprocessing and logical per-route costs are recorded separately.

O E2E TARGET-BY-CONSTRAINED-EMISSION FOLLOW-UP

We cross the two E2E supervision targets with two generation-time emission constraints on the predesignated 48 families: the first eight source-order families from each of six construction categories. All six noise/coordinate conditions and three training seeds are included, giving 864 returns per cell and 3456 primary returns. The six existing checkpoints after 2000 training steps are reused. These families were previously exposed in Cohort B; this is a cross-backbone follow-up. The complete predesignated panel defines this analysis.

The common native grammar retains operator arities, active variables, integer/named constants and float-span state. P excludes float-sign openers and admits `CONSTANT`; C excludes `CONSTANT` and admits float leaves. Native integer leaves and `pow2`/`pow3` remain in both. Each cell uses 64 proposals (one greedy, 63 sampled), maximum length 200, and the first eight valid, stable-unique prefixes in raw-draw order. Truncated and invalid draws are retained without resampling. P uses eight cold starts (R8); C uses its predicted start in place of R1 plus the shared R2–R8 (P8). Each start is capped at 200 actual forward calls, including gradient forwards, and 5 s after compilation/upload. The separately retained C/R8 control adds R1, so C can incur nine physical starts. Expressions use float64 scalar constants and support the native real-valued operators and named constants. Visible-MSE selection returns one formula; a no-winner result falls back to the visible mean.

The target ordering reverses between the constrained emissions at both high-precision thresholds. With $\Delta_e = (S - L)_e$, the primary interaction $I = \Delta_C - \Delta_P$ is -101.97 percentage points at high99, with a nine-comparison-adjusted interval of [-131.97,-72.08]. This difference of differences can exceed 100 percentage points in magnitude. It remains -88.31 [-120.89,-55.56] under shared random-start R8. The strict P simple-effect interval includes zero. For literal C, replacing R1 with the predicted start changes high99/high999/strict counts from 393/218/28 to 511/331/40. The other three cells' aggregate success counts are unchanged. This interaction includes both generation validity under constrained decoding and the contribution of initialization to literal/C refinement.

The raw pool contains 221,184 proposals, of which 109,788 are truncated and invalid; 14,468 parents reach refinement. Literal/P has 846 empty generations and 847 final fallbacks, while skeleton/C has 699 and 701, respectively. Both matched cells have zero empty generations or fallbacks. These counts locate the dominant mismatch at generation, before numerical fitting. All primary returned predictions are finite. Full mean FVU is 9.50×10^{13} for literal/P, 0.06923 for literal/C, 4.28×10^{18} for skeleton/P, and 1.89×10^{14} for skeleton/C; the complete medians appear in Table 25. Literal/P, skeleton/P, and skeleton/C retain 11, 1, and 61 returns above FVU 1,000, respectively. Literal/C has no return above FVU 10.

The evaluation accounts for all 121,772 physical starts and 15,967,957 expression forwards, including 52,032 forward-cap stops and 28 attempts with undefined or non-real symbolic expressions. Generation uses GPU inference; fitting uses the calibrated float64 CPU/CUDA implementations. Summed generation service time is 97,964.64 s after shared batches are deduplicated; fit-component time, including compilation, is 72,985.34 s, with 180.63 s of postprocessing. These component sums describe realized work under concurrent execution.

All saved visible predictions replay exactly. Native execution and the independent prefix evaluator agree on all 6,912 primary/secondary endpoint classifications. In two tasks, association as a/b versus $a(1/b)$ gives small numerical differences while preserving these classifications. Independent saved-array arithmetic and cluster-multiplicity resampling reproduce the summary counts and all 18 confidence intervals.

Table 25: All 864 returns per cell. Precision columns are counts at $R^2 \geq .99$, $R^2 \geq .999$, and FVU $\leq 2^{-23}$. Empty means no valid generated parent; fallback additionally includes failures after generation. All returned predictions are finite, but finite tail risk remains.

Target	Emit	High99	High999	Strict	Empty	Fallback	Mean FVU	Median FVU
Literal	P	0	0	0	846	847	9.50×10^{13}	1.005
Literal	C	511	331	40	0	0	0.06923	0.002597
Skeleton	P	370	184	16	0	0	4.28×10^{18}	0.02604
Skeleton	C	0	0	0	699	701	1.89×10^{14}	1.007

Table 26: Percentage-point contrasts with crossed three-seed $\times$ 48-family bootstrap intervals (10,000 resamples; all six conditions clustered). The primary nine comparisons and secondary nine comparisons use separate Bonferroni-9 correction. S/L denote skeleton/literal supervision. The full 18 contrasts and seedwise points are provided in the machine-readable export.

Contrast	High99	High999	Strict
Δ_P	+42.82 [+26.13, +59.26]	+21.30 [+10.42, +34.63]	+1.85 [+0.00, +4.86]
Δ_C	-59.14 [-73.61, -44.33]	-38.31 [-51.30, -25.69]	-4.63 [-9.26, -1.04]
$I = \Delta_C - \Delta_P$	-101.97 [-131.97, -72.08]	-59.61 [-84.03, -37.15]	-6.48 [-13.57, -1.39]
I, random starts	-88.31 [-120.89, -55.56]	-46.53 [-72.02, -24.05]	-5.09 [-11.92, -0.69]

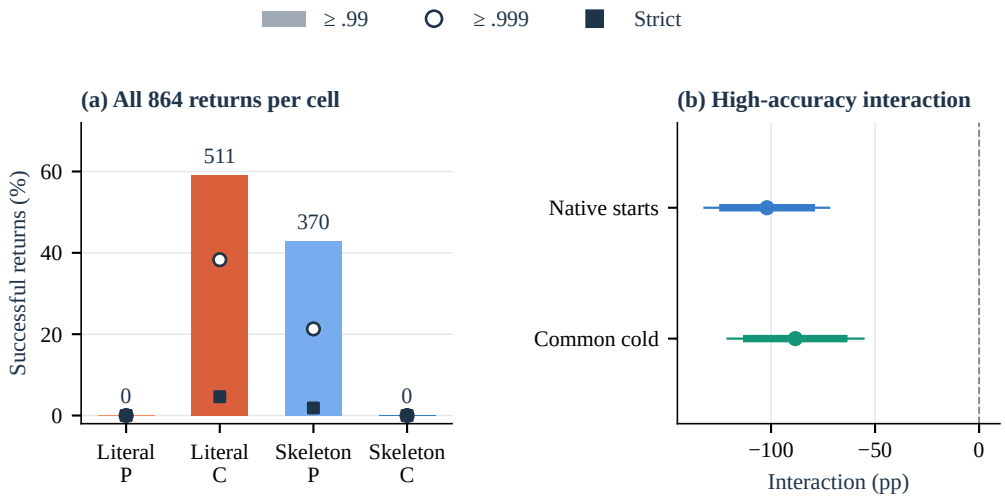


Figure 25: E2E target-by-constrained-emission follow-up on 48 exposed families (three seeds, six conditions; 864 returns per cell). (a) Bars show high-accuracy coverage, with success counts out of 864 printed above the bars. Open circles mark $R^2 \geq .999$ and squares mark strict success. (b) The high-accuracy interaction under native and shared random-start starts. Thick intervals are nominal 95%; thin intervals are Bonferroni-9 adjusted, separately for primary and secondary comparisons. E2E implements the two emission conditions through decoding constraints within its single native generation interface.

P INITIALIZER AVAILABILITY AND REFINEMENT STAGES

This analysis uses fixed eight-parent subsets of four existing 64-proposal pools: literal/role-masked SCALE-ERM training and placeholder/constants output, on the first 32 old conditions (six unequal exposed-family clusters, training seed 0). Literal constants supply a finite numeric initializer for 202/202 parameterized parents; role-masked constants supply 159/194, leaving 35/194 (18.0%) without a numeric vector. The 54/62 zero-FREE parents are separate.

The main geometry comparison matches the exact task and typed/shared-binding parameter family. There are 53 shared groups, of which 19 have comparable numeric vectors in both pools. The reference is the best accepted fit observed among random starts $R_1, \ldots, R_4$. Each group retains every original parent. Cached design ranks and covariance fields describe numerical conditioning; parameter identifiability is unassessed.

Table 27: Matched initializer geometry across the 19 comparable task/family groups. Differences are literal minus role-masked; *smaller* means lower error relative to the shared cold reference. Pooled distributions are descriptive because their parent-family compositions differ.

Error summary	Median difference	Literal smaller	Roles smaller	Tied
Absolute log-magnitude error	0.00	7	3	9
Sign-disagreement fraction	0.00	0	0	19
Vector-angle error (degrees)	0.00	6	3	10

No previously strict parent is lost from continuous to canonical or returned postprocessed stages, and S_1 misses no existing strict return. Postprocessing raises strict coverage for roles F from 1 to 3 conditions and for roles C from 0 to 1; the literal arms remain at 1. These caches measure refinement within the available candidate pools. Separating missing parent structure from optimization failure requires additional candidate-level evidence.

Q SENSITIVITY TO SUPERVISED-TOKEN WEIGHTING

The original objective first averages valid-token cross-entropy within each view, then averages the 16 views in an optimizer update. We replace it by a single mean over all valid supervised tokens in those views. If ℓ_j is the mean CE and n_j the supervised-token count in view j, the objectives are

$$L_{\text{view}} = \frac{1}{16}\sum_{j=1}^{16} \ell_j, \qquad L_{\text{token}} = \frac{\sum_{j=1}^{16} n_j \ell_j}{\sum_{j=1}^{16} n_j}.$$

Both literal and role targets receive the same normalization rule. Six new 20M continuations retain the original weights, observations, paired-scale views, target serialization, prompts, native label masks, family-index stream, optimizer, schedule, and 2,000 updates. Four four-view forwards accumulate gradients using the common 16-view token denominator; all evaluations use the fixed final checkpoint. A direct native CE reduction and the streamed weighted reduction agree in the implementation check.

Each final checkpoint uses both output requests on cohort B families 0–31, all three noise levels and two views. This exposed-family sensitivity panel contains 2,304 new returns and reuses 2,304 matched original returns, each with 64 proposals and the unchanged native refinement/selection pipeline.

Global token averaging retains the supervision ranking reversal at $R^2 \geq .99$: roles exceed literals under F (343 versus 307), while literals exceed roles under C (355 versus 341). The interaction is -8.68 points, compared with -8.51 under per-view averaging on the same panel. It is negative in all three seeds under both rules. The global-token interaction has a descriptive nominal interval $[-17.36, -1.04]$; the matched change in interaction is -0.17 points, with adjusted interval $[-14.29, 14.58]$. This control measures sensitivity to loss weighting while retaining each format's prompts and sequences.

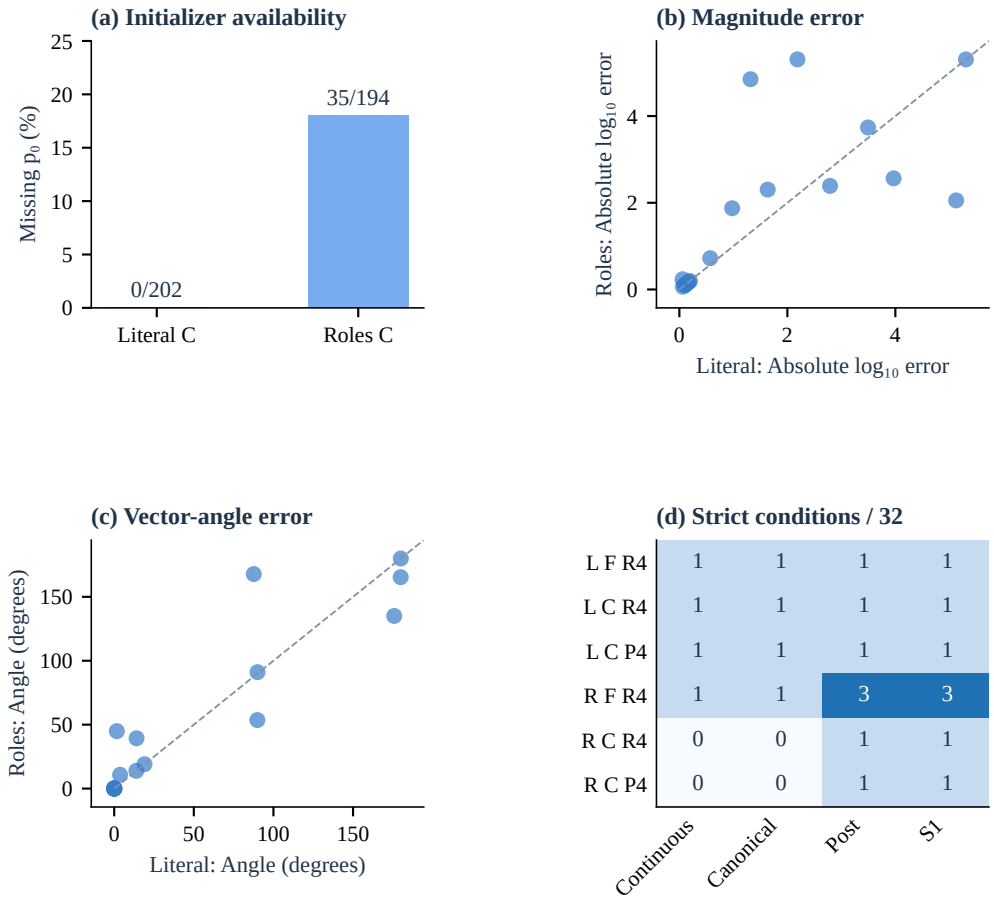


Figure 26: Cache-only mechanism audit. (a) Missing numeric initializers among parameterized constants parents. (b,c) All 19 exact task/family pairs; coincident points overlap. Magnitude summaries use jointly nonzero coordinates, with zero-ratio events retained separately and no epsilon floor. (d) Conditions with at least one strict parent in each recorded pool, followed by the frozen visible S_1 return; strict means audit FVU $\leq 2^{-23}$. F/C denote placeholder/constants output. The stage result applies to this 32-condition, fixed-eight-parent mechanism pool.

Table 28: Loss-normalization sensitivity. Each cell contains 576 conditions across three training seeds. Strict denotes FVU $\leq 2^{-23}$; NF counts nonfinite-return conditions.

CE mean	Target	Output	$\geq .99$	$\geq .999$	Strict	NF	Full mean FVU
Per view	Literal	F	309	184	47	1	∞
Per view	Literal	C	359	224	56	0	0.0407
Per view	Roles	F	335	226	57	0	0.0425
Per view	Roles	C	336	218	53	0	0.0483
Per token	Literal	F	307	190	48	0	0.0485
Per token	Literal	C	355	241	62	0	0.0370
Per token	Roles	F	343	223	53	0	0.0477
Per token	Roles	C	341	214	51	0	0.0414

Table 29: Global-token minus per-view differences in percentage points. Paired crossed resampling uses 32 family blocks and three training seeds, retaining all six within-family conditions. Bonferroni-15 covers four cell differences and their interaction change at three endpoints.

Comparison	Endpoint	Δ	Nominal 95%	Adjusted interval
Literal F	$R^2 \geq .99$	-0.35	$[-5.21, 4.34]$	$[-8.16, 6.48]$
Literal C	$R^2 \geq .99$	-0.69	$[-6.60, 3.65]$	$[-10.30, 5.73]$
Roles F	$R^2 \geq .99$	$+1.39$	$[-5.73, 8.33]$	$[-9.14, 13.19]$
Roles C	$R^2 \geq .99$	$+0.87$	$[-5.04, 7.47]$	$[-8.33, 11.98]$
Interaction change	$R^2 \geq .99$	-0.17	$[-9.38, 9.55]$	$[-14.29, 14.58]$
Literal F	$R^2 \geq .999$	$+1.04$	$[-2.08, 4.69]$	$[-3.82, 6.77]$
Literal C	$R^2 \geq .999$	$+2.95$	$[-2.60, 8.33]$	$[-6.25, 11.63]$
Roles F	$R^2 \geq .999$	-0.52	$[-7.99, 6.42]$	$[-12.50, 9.90]$
Roles C	$R^2 \geq .999$	-0.69	$[-6.60, 6.08]$	$[-9.90, 9.90]$
Interaction change	$R^2 \geq .999$	-2.08	$[-11.63, 6.77]$	$[-18.58, 11.23]$
Literal F	Strict	$+0.17$	$[-2.78, 3.12]$	$[-4.17, 4.69]$
Literal C	Strict	$+1.04$	$[-3.13, 5.38]$	$[-5.73, 8.39]$
Roles F	Strict	-0.69	$[-5.03, 3.65]$	$[-7.18, 6.77]$
Roles C	Strict	-0.35	$[-2.95, 2.26]$	$[-4.69, 4.05]$
Interaction change	Strict	-0.52	$[-7.29, 4.69]$	$[-11.46, 7.12]$

R VIEWS PER FAMILY AT FIXED TOTAL EXPOSURE

The pooled high-accuracy supervision preferences persist at two, four and eight views per family: roles favor placeholder emission, whereas literal targets favor constant emission (Table 30). The corresponding interactions are -6.42, -8.51 and -4.86 percentage points. All nine training-seed-specific interaction estimates are negative. On this exposed 32-family panel, the pointwise 95% intervals are $[-14.76, 1.39]$, $[-16.15, -1.56]$ and $[-11.63, 1.56]$; all three intervals include zero after the nine-test adjustment across q and precision thresholds. The 30 adjusted comparisons of $q = 2$ or $q = 8$ against $q = 4$ also include zero.

We train 12 new 20M checkpoints: $q \in \{2, 8\}$, literal or role targets, and three training seeds. Every run uses 2,000 updates, 16 total views per update, 128 observations and the original optimizer and per-view loss normalization. The two-view condition retains the original identity and first scale view. The eight-view condition retains all four original views and adds four independent scales from the same $[-6, 6]$ log-scale range. Original serialized targets are retained exactly. Evaluation uses the first 32 families of cohort B, all three noise levels and both coordinate views, with 64 proposals per return. The 4,608 new returns are compared with the same 2,304 archived $q = 4$ returns.

At fixed exposure, $q = 2$ uses eight family groups per update and $q = 8$ uses two, totaling 16,000 and 4,000 family presentations. This varies views per family together with family diversity and the fraction of identity views. Both placeholder configurations have higher observed coverage at $q = 2$ than at $q = 4$, while constant-emission changes differ by target representation. The literal/placeholder cells retain non-finite returns at every q.

S MIXING OUTPUT REQUESTS WITHIN A BALANCED CHECKPOINT

A balanced checkpoint supports both interfaces, raising the deployment question of how to divide a fixed proposal budget. We compare two independent 32-draw placeholder searches (F+F), two constant searches (C+C), and one of each (F+C). Each policy receives 64 proposals. We merge the raw expression beams with stable native deduplication, then run the full native numerical fitting, typed-literal and discrete refinement, and visible MDL selector. We merge candidate pools before fitting.

The fixed exposed panel contains cohort B family indices $0, 8, \ldots, 248$, three noise levels, two views, and all three 20M balanced checkpoints. The 32 families belong to generator categories 0, 2, and 4. Each policy therefore returns 576 formulas. Four raw pools per condition are shared across comparisons, giving 73,728 physically generated proposals and 110,592 logical proposals over the three policies. Each policy undergoes its own full refinement and selection.

Table 30: View-count sensitivity with 576 returns per row: 32 exposed families, six noise/coordinate conditions and three training seeds. F/C denote placeholder and constant emission; .99 and .999 are R^2 thresholds. Strict numerical success uses FVU $\leq 2^{-23}$. NF and the full-denominator mean retain all non-finite predictions. Description lengths use the native returned expression. All new training runs use the same local GPU. For new inference, seed 0 runs locally and seeds 1–2 run on the server.

q	Target	Output	.99	.999	Strict	NF	Mean FVU	Median bits
2	Literal	F	328	188	53	2	∞	62.17
2	Roles	F	357	235	64	0	0.04436	62.17
2	Literal	C	354	228	50	0	0.04139	59.99
2	Roles	C	346	239	55	0	0.04764	65.67
4	Literal	F	309	184	47	1	∞	61.18
4	Roles	F	335	226	57	0	0.04250	63.71
4	Literal	C	359	224	56	0	0.04069	61.67
4	Roles	C	336	218	53	0	0.04832	65.99
8	Literal	F	305	190	47	1	∞	61.95
8	Roles	F	311	186	49	0	0.04944	62.82
8	Literal	C	351	222	60	0	0.03800	63.17
8	Roles	C	329	210	55	0	0.04689	64.55

Table 31: Balanced-checkpoint deployment at 64 proposals per return. All policies have zero empty or nonfinite returns. Counts use the full 576-condition denominator. Refinement CPU seconds sum the recorded work under the fixed proposal budget.

Policy	$R^2 \geq .99$	$\geq .999$	Strict	Mean FVU	Median bits	Refine CPU s
F+F	372	278	66	0.0392	58.79	1858.6
C+C	400	279	60	0.0329	59.90	2118.7
F+C	385	275	66	0.0353	57.96	1983.1

F+C adds 13 high-accuracy returns over F+F but loses 15 relative to C+C. It matches F+F's strict count and exceeds C+C by six. All nine adjusted intervals include zero. The balanced checkpoint supports both requests. Under this budget, the paired interval for mixed versus repeated-constant search includes zero.

Table 32: Paired policy differences in percentage points, with crossed 32-family/three-seed intervals. Adjustment covers three policy pairs and three endpoints. FF/CC/FC abbreviate the policies above.

Comparison	Endpoint	Δ	Nominal 95%	Bonferroni-9
FC − FF	$\geq .99$	$+2.26$	$[-2.08, 6.60]$	$[-4.17, 8.68]$
FC − FF	$\geq .999$	-0.52	$[-3.65, 2.60]$	$[-5.59, 4.51]$
FC − FF	Strict	$+0.00$	$[-2.60, 2.60]$	$[-4.17, 3.86]$
FC − CC	$\geq .99$	-2.60	$[-5.73, 0.35]$	$[-7.29, 1.56]$
FC − CC	$\geq .999$	-0.69	$[-3.82, 2.43]$	$[-5.21, 3.82]$
FC − CC	Strict	$+1.04$	$[-1.22, 3.12]$	$[-2.60, 4.17]$
CC − FF	$\geq .99$	$+4.86$	$[-0.00, 9.72]$	$[-2.60, 12.15]$
CC − FF	$\geq .999$	$+0.17$	$[-3.99, 4.51]$	$[-6.25, 6.60]$
CC − FF	Strict	-1.04	$[-3.65, 1.74]$	$[-5.21, 3.12]$

Paired interface complementarity at 64 proposals. Pairing the 576 saved returns from each policy shows 42 CC-only and 14 FF-only high-accuracy successes (Table 33). FC recovers 30 and seven of these, respectively, but loses ten conditions where both single-format policies succeed. It adds no high-accuracy success among their 162 joint failures. Its selected parents originate from F in 283 conditions and C in 293.

At the strict endpoint, FC gains and loses nine successes against FF; against CC it gains 13 and loses seven. All policies have zero nonfinite or empty returns and zero FVU > 10 cases. Their 99th-percentile FVUs are 0.33976, 0.33597, and 0.33597 for FF, CC, and FC. Each policy executes

Table 33: Paired $R^2 \geq .99$ transitions at 64 proposals, with all 576 conditions per comparison. These exposed 32 families cover three generator categories; the transitions quantify overlap between the success sets of the two output modes.

A	B	Both fail	A only	B only	Both succeed
FC	FF	174	30	17	355
FC	CC	169	7	22	378
CC	FF	162	42	14	358

576 complete native refinements, with 30,728, 32,921, and 32,332 retained parents, respectively. Table 31 reports measured CPU work; individual optimizer-call and forward-call counts were not recorded.

T PUBLIC TRANSFER OF BOTH OUTPUT INTERFACES

Table 34: All public target–interface cells and training seeds. .99 and .999 count the respective R^2 thresholds. FVU summaries retain every return, including the mean fallbacks. Strict and nonfinite counts are zero in every row. Bits are the median recorded code length among native returns; fallbacks have no symbolic code length. Pooled rows contain 216 returns.

Target/output	Seed	n	.99	.999	Mean FVU	Median FVU	Native	Bits
Literal F	0	72	15	5	0.7931	0.3145	69	37.59
Literal F	1	72	16	5	0.3764	0.3227	69	38.71
Literal F	2	72	19	6	0.3310	0.2869	69	43.17
Literal F	pooled	216	50	16	0.5002	0.3137	207	39.39
Literal C	0	72	16	4	0.3149	0.2603	69	38.71
Literal C	1	72	17	4	0.3419	0.2893	69	37.26
Literal C	2	72	19	7	0.3246	0.2634	69	45.30
Literal C	pooled	216	52	15	0.3271	0.2746	207	40.39
Roles F	0	72	16	5	0.3588	0.2364	69	39.46
Roles F	1	72	16	4	0.3174	0.2154	69	39.68
Roles F	2	72	17	4	0.3235	0.2592	69	45.00
Roles F	pooled	216	49	13	0.3332	0.2500	207	42.30
Roles C	0	72	15	4	0.3818	0.3120	69	42.42
Roles C	1	72	16	7	0.3252	0.2855	69	39.46
Roles C	2	72	19	7	0.3460	0.2993	69	39.14
Roles C	pooled	216	50	18	0.3510	0.3069	207	39.46
50/50 F	0	72	17	5	0.4162	0.2897	69	45.60
50/50 F	1	72	19	6	0.3137	0.2207	69	45.59
50/50 F	2	72	17	4	0.3508	0.3083	69	40.08
50/50 F	pooled	216	53	15	0.3602	0.3014	207	44.11
50/50 C	0	72	19	5	0.3222	0.2595	69	40.70
50/50 C	1	72	17	4	0.3808	0.2951	69	39.39
50/50 C	2	72	17	6	0.3217	0.2911	69	39.39
50/50 C	pooled	216	53	15	0.3416	0.2828	207	39.39

Balanced-format continuation records 53 high-accuracy returns out of 216 with each output interface. Its three seed counts are 17/19/17 under placeholders and 19/17/17 under constants. Every balanced checkpoint therefore exceeds the observed high-accuracy counts of the evaluated Operon (12/72), PySR (10/72), and TIR (7/72) configurations. The median-error ranking differs: Operon and TIR obtain 0.2262 and 0.2268, compared with balanced F/C's 0.3014/0.2828. Figure 6 displays coverage and both complete mean and median risk.

Matched public queries. The six cells cross literal, role, and balanced scale-trained 20M checkpoints with placeholder (F) and constants (C) requests. Each uses all three training seeds, all 24 already exposed datasets, and their three fixed official splits. All neural queries use 1,024 proposals, the original visible training rows, and the native fitting, refinement, and selection policy. Each

checkpoint retains 69 native returns and three predefined mean fallbacks for unsupported inputs. Existing placeholder results and seed-0 constants results are reused; the two additional specialist constants seeds and balanced outputs complete the matrix. All 1,296 returns have finite predictions and none reaches strict success. The baseline configurations and their original budgets are specified in Appendix I. Fixed baselines contribute one fitted set of 72 conditions each, shared when comparing training seeds.

Paired comparisons. We jointly resample the 24 dataset blocks and three training seeds, keeping all three data splits together, for 10,000 bootstrap draws. Tables below report percentage-point differences with nominal 95% intervals and Bonferroni-adjusted intervals within the nine interface, twelve balanced, and eight baseline comparisons. Their shared resampling indices preserve pairing across cells. The high-accuracy interaction is -0.46 points (nominal $[-11.11, 8.80]$; adjusted $[-15.28, 13.89]$), with seed effects $-2.78, -1.39, +2.78$. The adjusted public interaction interval includes both directions. All adjusted balanced and baseline intervals also contain zero. Strict bootstrap intervals collapse to zero because every observed strict count is zero. The observed count leads refer to these specific system configurations and their respective budgets.

Table 35: Public supervision–interface comparisons, with adjustment for nine contrasts. L, R, M denote literal, roles, and balanced training; F/C denote output interfaces. All differences and intervals are percentage points.

Contrast	Endpoint	Difference	Nominal 95%	Adjusted 95%
$R_F - L_F$	$R^2 \geq .99$	-0.46	$[-7.87, 6.94]$	$[-12.04, 10.19]$
$R_C - L_C$	$R^2 \geq .99$	-0.93	$[-5.56, 3.24]$	$[-7.51, 6.48]$
$(R_C - L_C) - (R_F - L_F)$	$R^2 \geq .99$	-0.46	$[-11.11, 8.80]$	$[-15.28, 13.89]$
$R_F - L_F$	$R^2 \geq .999$	-1.39	$[-5.56, 2.78]$	$[-7.41, 4.17]$
$R_C - L_C$	$R^2 \geq .999$	$+1.39$	$[-1.85, 6.48]$	$[-3.70, 9.72]$
$(R_C - L_C) - (R_F - L_F)$	$R^2 \geq .999$	$+2.78$	$[-3.24, 9.72]$	$[-6.94, 13.89]$
$R_F - L_F$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$
$R_C - L_C$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$
$(R_C - L_C) - (R_F - L_F)$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$

Table 36: Balanced training against each specialist at the same public output interface; adjustment for twelve contrasts. Units and notation match the preceding table.

Contrast	Endpoint	Difference	Nominal 95%	Adjusted 95%
$M_F - L_F$	$R^2 \geq .99$	$+1.39$	$[-5.56, 8.33]$	$[-9.72, 13.89]$
$M_F - R_F$	$R^2 \geq .99$	$+1.85$	$[-4.63, 9.72]$	$[-9.34, 15.28]$
$M_C - L_C$	$R^2 \geq .99$	$+0.46$	$[-4.63, 6.02]$	$[-6.94, 9.72]$
$M_C - R_C$	$R^2 \geq .99$	$+1.39$	$[-4.17, 7.87]$	$[-6.94, 11.57]$
$M_F - L_F$	$R^2 \geq .999$	-0.46	$[-4.63, 4.63]$	$[-6.94, 8.33]$
$M_F - R_F$	$R^2 \geq .999$	$+0.93$	$[0.00, 3.70]$	$[0.00, 6.48]$
$M_C - L_C$	$R^2 \geq .999$	$+0.00$	$[-2.78, 2.78]$	$[-5.56, 4.17]$
$M_C - R_C$	$R^2 \geq .999$	-1.39	$[-6.48, 2.79]$	$[-9.72, 7.41]$
$M_F - L_F$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$
$M_F - R_F$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$
$M_C - L_C$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$
$M_C - R_C$	Strict	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$

Table 37: Public high-accuracy differences from the fixed classical baselines; adjustment for eight contrasts. Baseline predictions are reused across neural training seeds, not treated as additional independent baseline fits.

Contrast	Endpoint	Difference	Nominal 95%	Adjusted 95%
Roles F − PySR	$R^2 \geq .99$	$+8.80$	$[-4.17, 23.61]$	$[-8.80, 29.98]$
Roles F − Operon	$R^2 \geq .99$	$+6.02$	$[-9.72, 22.70]$	$[-15.63, 29.17]$
Literal C − PySR	$R^2 \geq .99$	$+10.19$	$[-0.46, 23.61]$	$[-3.70, 30.56]$
Literal C − Operon	$R^2 \geq .99$	$+7.41$	$[-5.56, 22.69]$	$[-9.72, 29.63]$
50/50 F − PySR	$R^2 \geq .99$	$+10.65$	$[-0.46, 24.54]$	$[-4.17, 30.56]$
50/50 F − Operon	$R^2 \geq .99$	$+7.87$	$[-5.09, 23.16]$	$[-9.72, 30.09]$
50/50 C − PySR	$R^2 \geq .99$	$+10.65$	$[-0.46, 24.09]$	$[-4.17, 30.56]$
50/50 C − Operon	$R^2 \geq .99$	$+7.87$	$[-5.57, 23.61]$	$[-10.53, 30.90]$

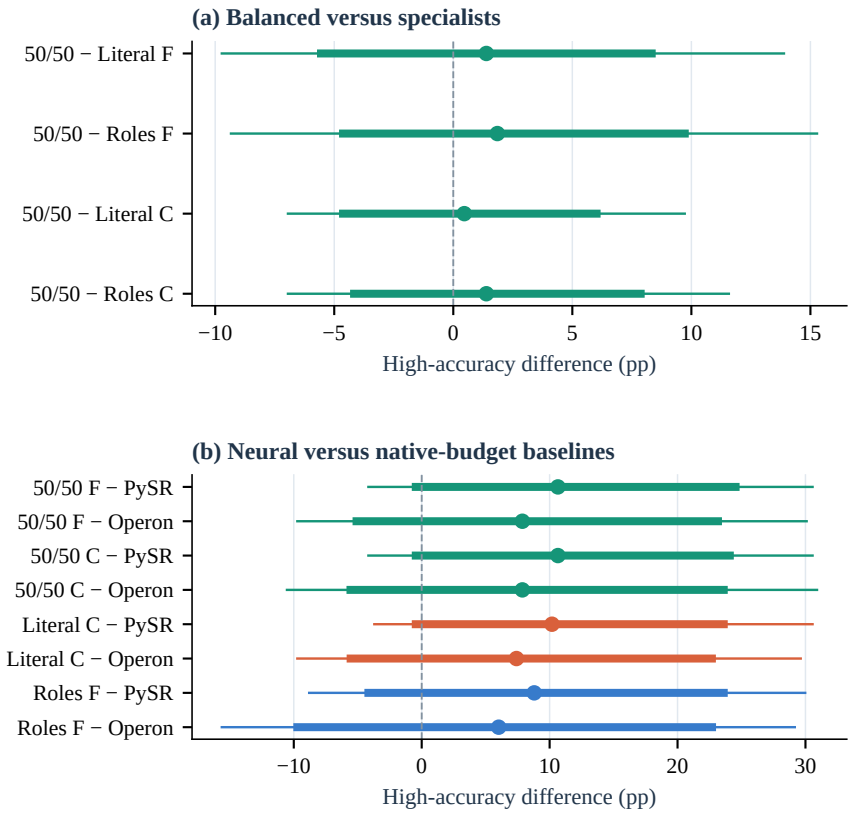


Figure 27: **Public high-accuracy effects with their uncertainty.** (a) Balanced training minus each specialist under the same output request. (b) Neural configurations minus native-budget PySR and Operon. Thick intervals are nominal 95%; thin intervals reproduce the original adjusted intervals in the full public comparison tables. Positive point estimates are observed coverage gains; each adjusted interval includes both positive and negative differences.

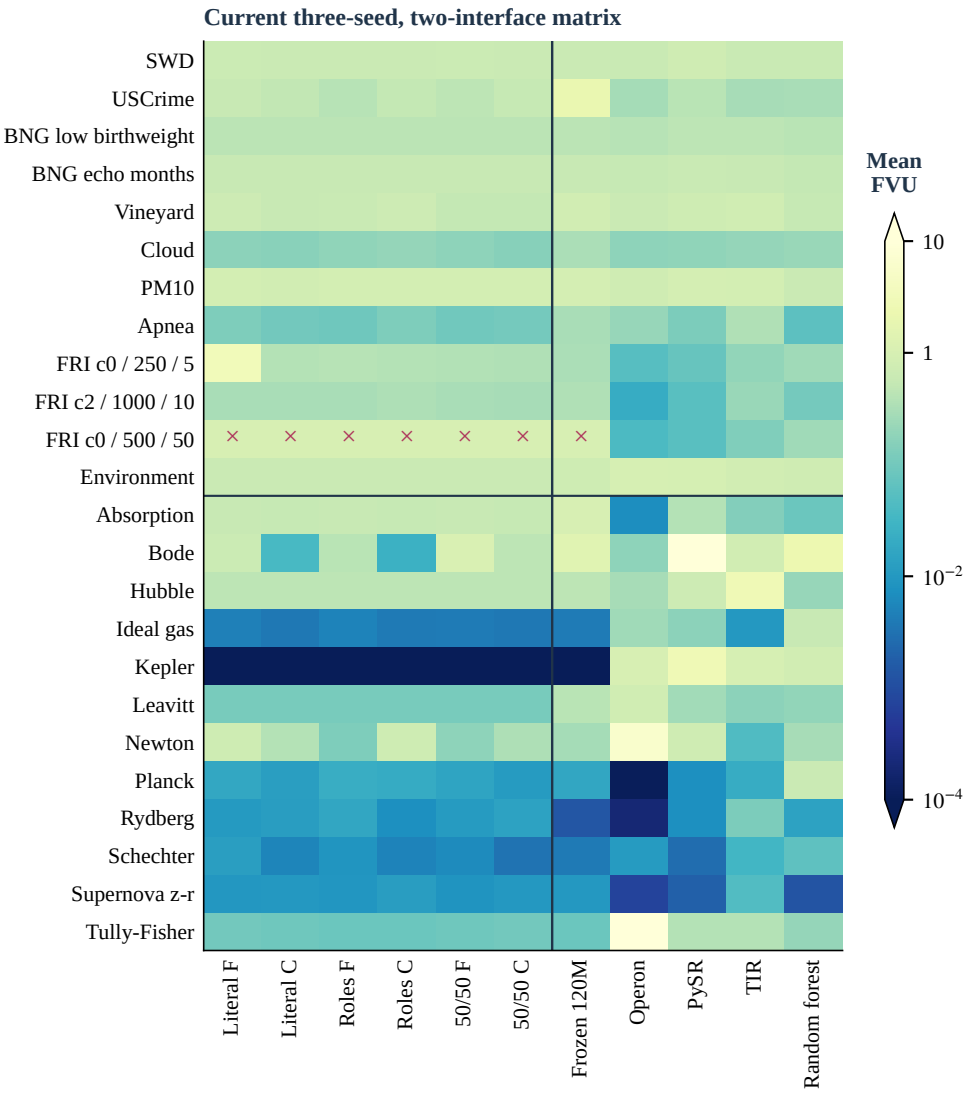


Figure 28: **The complete public matrix by dataset.** The six current neural cells each average nine split–training-seed returns per dataset; each fixed baseline averages its original three splits. Colors display mean FVU on a logarithmic scale clipped to $[10^{-4}, 10]$; the full numerical values remain in the accompanying table data and risk summaries. A cross marks a cell containing a documented mean fallback. Dataset order follows the fixed source identifiers. This current matrix is distinct from the historical seed-0 heatmap.

U A SEPARATE 120M SHORT-BUDGET PUBLIC EVALUATION

The short-budget experiment evaluates three balanced 120M checkpoints, a frozen 120M checkpoint, PySR, and Operon on the same 24 exposed datasets and three official splits. The protocol was registered after earlier public-data exposure. Neural queries use the constants request (C), 256 proposals, at most 1,024 training rows, and visible-data RMS conditioning. PySR and Operon each receive a native 120 s budget with four threads under the same row cap and coordinates. This yields 288 neural and 144 classical returns. The 20M public matrix in §T uses 1,024 proposals and all available training rows in raw coordinates; its results remain a separate comparison.

Table 38: All 432 short-budget returns; every neural model uses the constants request (C). The pooled balanced row combines its three training seeds; other methods have one return per official split. NF counts nonfinite predictions. Every mean includes all returns, including the fixed mean fallbacks. Strict is FVU $\leq 2^{-23}$.

Model	n	High99	High999	Strict	NF	Median FVU	Mean FVU
Balanced / seed 0	72	16	6	0	0	0.2505	0.3496
Balanced / seed 1	72	13	4	0	0	0.3241	0.645
Balanced / seed 2	72	10	5	0	0	0.3295	0.395
Frozen 120M	72	9	0	0	1	0.4028	∞
Operon / 120 s	72	12	9	0	0	0.2293	25.65
PySR / 120 s	72	10	6	0	1	0.2452	∞
Balanced / pooled	216	39	15	0	0	0.3102	0.4632

The balanced checkpoints yield 16, 13, and 10 high-accuracy returns per 72 splits, compared with 9 for frozen 120M, 10 for PySR, and 12 for Operon. At $R^2 \geq .999$, balanced training yields 6, 4, and 5, while the three comparators yield 0, 6, and 9. Median and mean risks favor different methods (Table 38). Three predefined mean fallbacks per neural checkpoint preserve unsupported input-dimensionality cases in the denominator. One frozen-120M return and one PySR return are nonfinite, giving infinite full means; all balanced and Operon predictions are finite.

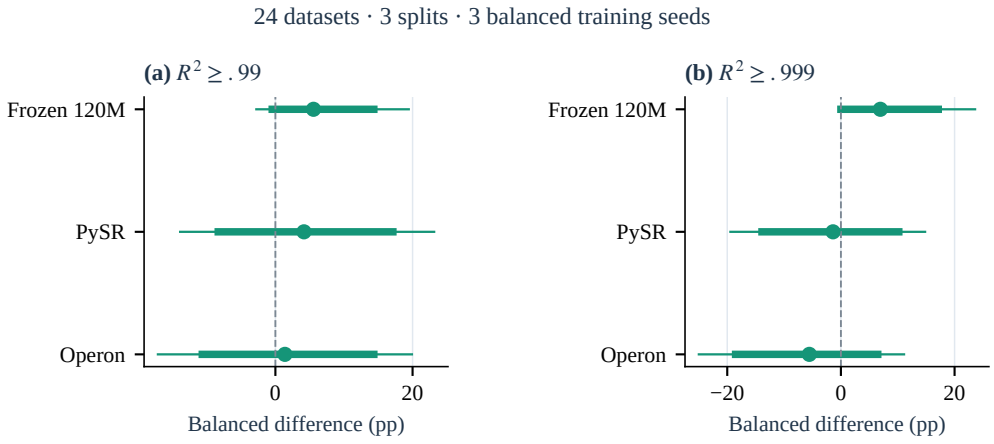


Figure 29: Balanced 120M minus each short-budget comparator on the fixed public panel. Thick intervals are nominal 95%; thin intervals adjust the nine contrasts in Table 39. Resampling jointly accounts for datasets and balanced training seeds, with all three splits of a dataset clustered. Classical baseline runs are not replicated to create extra trials.

All high-accuracy adjusted intervals include zero. The balanced–frozen $R^2 \geq .999$ interval touches zero, and all strict counts are zero. The comparisons report observed accuracy and resource use under this protocol, with adjusted intervals that include zero for every high-accuracy difference.

Table 39: All nine short-budget contrasts, in percentage points. Each compares pooled balanced performance with a fixed comparator. The strict intervals collapse to zero because every observed strict count is zero.

Balanced minus	Endpoint	Difference	Nominal 95%	Adjusted 95%
Frozen 120M	$R^2 \geq .99$	+5.56	[-0.46, +14.35]	[-2.78, +19.44]
PySR / 120 s	$R^2 \geq .99$	+4.17	[-8.33, +17.13]	[-13.89, +23.15]
Operon / 120 s	$R^2 \geq .99$	+1.39	[-10.66, +14.35]	[-17.13, +19.91]
Frozen 120M	$R^2 \geq .999$	+6.94	[+0.00, +17.13]	[+0.00, +23.61]
PySR / 120 s	$R^2 \geq .999$	-1.39	[-13.89, +10.19]	[-19.44, +14.81]
Operon / 120 s	$R^2 \geq .999$	-5.56	[-18.53, +6.48]	[-25.00, +11.11]
Frozen 120M	Strict	+0.00	[+0.00, +0.00]	[+0.00, +0.00]
PySR / 120 s	Strict	+0.00	[+0.00, +0.00]	[+0.00, +0.00]
Operon / 120 s	Strict	+0.00	[+0.00, +0.00]	[+0.00, +0.00]

Native author-environment predictions define the reported performance. Independent saved-array arithmetic and bootstrap-weight checks pass for all 432 returns. Mathematical exports provide an additional diagnostic for the 144 classical returns: all threshold classifications and finite/nonfinite kinds agree, while numerical export discrepancies remain recorded. The complete native predictions and discrepancy records are retained.

Summed service times for the three balanced checkpoints are 1,429.16, 1,384.68, and 1,457.76 s; frozen 120M uses 1,065.06 s, Operon 8,647.95 s, and PySR 8,668.21 s. These totals account for generation, fitting, and postprocessing across concurrently executed tasks. These service times measure resource use during concurrent server execution.

V RMS TRAINING AND A BOUNDED PUBLIC TRANSFER CONFIGURATION

RMS-canonical training reaches 15 and 16 high-accuracy returns out of 72 public splits under placeholder and constant emission, respectively (Table 40). Both counts exceed scale-augmented training by two returns. The paired dataset-block 95% intervals for these differences are $[-2.78, 9.72]$ and $[0.00, 6.94]$ percentage points. The coverage increase coexists with larger mean FVU and longer recorded service times. All four configurations have zero strict numerical successes.

This exploratory transfer check uses the two existing literal 120M, seed-0 checkpoints, selected after the 32-family scale panel. Each configuration uses 256 proposals and at most 1,024 training observations. A common deterministic sample without replacement depends only on the task seed and row count. Input and target RMS values use the full training set, without centering; the native solver, initialization, typed rules and selector are unchanged. Predictions are transported back to the original coordinates and evaluated on every official test point. These settings differ from the 1,024-proposal, full-training-set public runs. The three official splits represent data splits, with one training seed for each checkpoint.

Table 40: Public RMS transfer, 72 returns per row: 24 datasets and three official splits. All rows include the same three 50-variable mean fallbacks. .99 and .999 denote the corresponding R^2 thresholds; strict means FVU $\leq 2^{-23}$. Mean, median and maximum are full-denominator test FVU. NF counts returns with a non-finite test prediction. Time is median recorded generation-plus-refinement service time, in seconds; the zero-service unsupported fallbacks remain in this median. F and C denote placeholder and constant emission.

Training	Output	.99	.999	Strict	NF	Mean	Median	Max.	Time (s)
Scale	F	13	6	0	0	0.7048	0.3627	16.74	15.10
Scale	C	14	5	0	0	1.3026	0.2931	66.51	18.20
RMS	F	15	7	0	0	1.4655	0.2752	42.90	45.66
RMS	C	16	6	0	0	1.5468	0.3132	66.51	57.88

A train-only variable interface. The only dataset beyond the model's 17-variable interface has 50 inputs. We evaluate three fixed 17-column selections on its three official splits: the first 17 columns, train-only mutual-information (MI) ranking, and a seeded random subset. Train-only variable prescreening is motivated by prior symbolic regression work (Ye & Li, 2025); this control uses MI with three neighbors. MI ties are resolved by original column index, and every selected set is passed to the model in ascending original index. All 375 training observations and the same 256-proposal budget are used. RMS factors come from the complete original training matrix; the returned expression is mapped to its original columns.

The first-17 and MI selections produce mean FVU between 0.295 and 0.446, compared with 0.935–1.031 for random selection (Table 41). All 36 returns are finite native predictions; none reaches either high-accuracy threshold or strict numerical success. This separate extension retains the 12 unsupported fallbacks in the main 288-return comparison.

Table 41: Mean test FVU for the separate 50-to-17-variable extension. Each cell contains all three official splits of one dataset. The two checkpoints and two output interfaces match Table 40. There are zero high-accuracy, strict and non-finite returns in each cell.

Training	Output	First 17	MI top 17	Random 17
Scale	F	0.4146	0.4464	0.9546
Scale	C	0.4415	0.4037	0.9346
RMS	F	0.3500	0.2948	1.0309
RMS	C	0.3109	0.3656	0.9902

W Coordinate selection across numerical scales

On all 720 conditions of the 120M SRSD catalogue, the three selectors act on the same pair: the native top return from each 1,024-draw raw/RMS search. Native, stored-visible-error, and transported selection give high-accuracy/strict counts of $585/284$, $585/252$, and $552/289$, respectively. We relate these differences to the task's RMS scales and the description-length increment incurred by transport.

We additionally evaluate $S_\lambda = \lambda S_{\text{native}} + (1-\lambda) S_{\text{transport}}$, for three jointly fixed weights $\lambda \in \{.25, .50, .75\}$, on the same finite-native-score pool with exact ties to raw. All three weights are exploratory results on this exposed catalogue.

Table 42: All policies on the same 720 conditions. High precision is $R^2 \geq .99$; strict is FVU $\leq 2^{-23}$. Tail counts include nonfinite predictions. Bits price the full original-coordinate description. Every full-denominator mean FVU is infinite; medians are shown.

Policy	High	Strict	FVU > 10	NF	RMS chosen	Median FVU	Median bits
Raw only	499	244	20	6	0	2.51×10^{-5}	54.0
RMS only	541	207	24	7	720	2.99×10^{-5}	371.2
Native (N)	585	284	15	5	417	2.62×10^{-6}	185.9
Error only (E)	585	252	17	6	349	3.02×10^{-6}	120.7
Transport (T)	552	289	15	3	79	2.02×10^{-6}	56.2
$\lambda = .25$	559	289	14	3	90	1.96×10^{-6}	56.6
$\lambda = .50$	566	290	13	3	104	1.95×10^{-6}	56.6
$\lambda = .75$	577	290	11	3	145	1.90×10^{-6}	58.6

Native selection adds 34 high-precision successes and loses one versus transport; strict adds 15 and loses 20. Of the net 33 high-precision gains, 32 occur at scale > 4 (Table 43), a within-cache association across physical problems.

The fixed code-length bins have little internal resolution: 718/720 RMS transports add more than 64 bits. The empty middle stratum is retained, and every interpolation weight is reported.

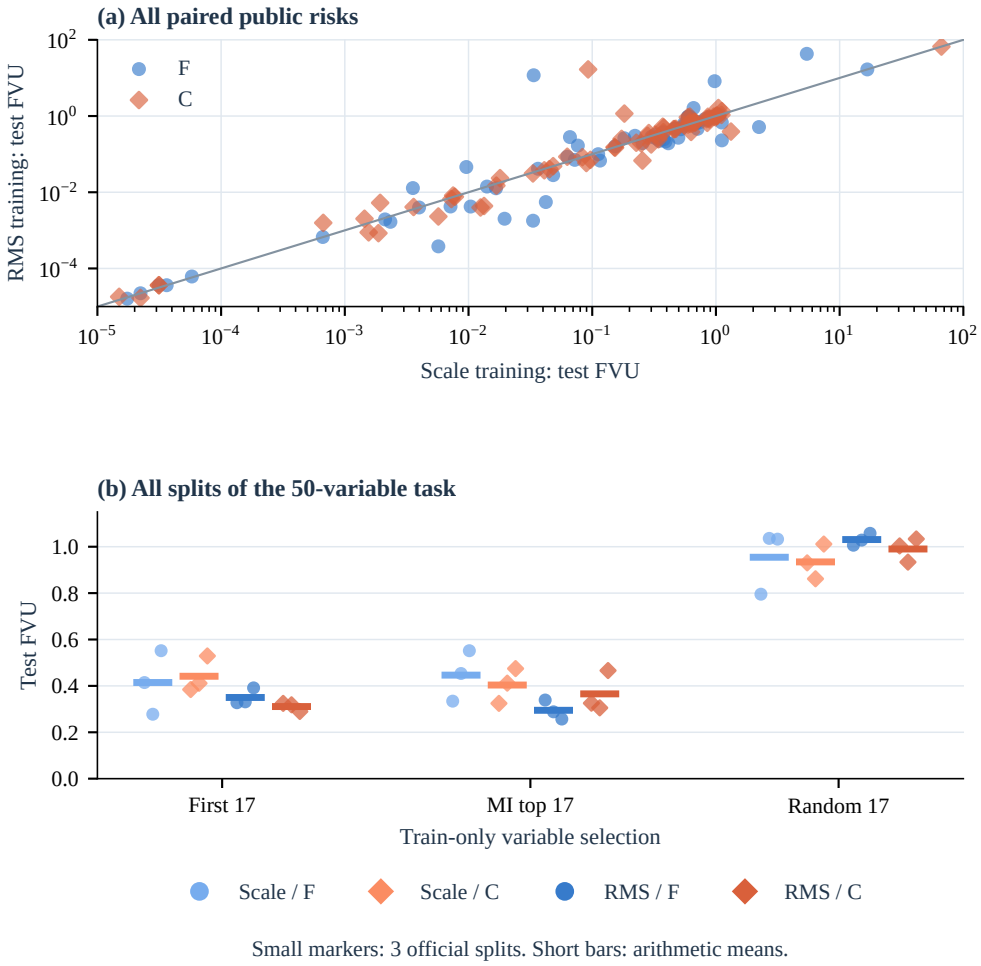


Figure 30: Complete public risks under the two training schemes and a separate variable-selection control. (a) Each point pairs one of the 72 official splits between scale and RMS-canonical training; both output interfaces are shown. The diagonal marks equal FVU. (b) Small markers show the three official splits for each checkpoint/interface cell; short horizontal bars show their arithmetic means. Horizontal offsets separate configurations and splits within each variable-selection method. All observed risks are retained in both panels.

Table 43: Fixed task strata, common to every policy. Scale is $\max_j |\log_{10} s_j|$, including input and target RMS scales. Code-length increment is the RMS transported-description bits minus its native-description bits. Cells report high/strict counts; the prespecified empty bit stratum is retained.

Stratum	n	Native	Error only	Transport
Scale $[0, 2]$	114	92/62	91/49	91/62
Scale $(2, 4]$	43	40/23	40/18	40/24
Scale > 4	563	453/199	454/185	421/203
Bits ≤ 0	2	2/2	2/2	2/2
Bits $(0, 64]$	0	–	–	–
Bits > 64	718	583/282	583/250	550/287

Table 44: All nine planned contrasts at both endpoints. Entries are percentage-point differences with Bonferroni-18 intervals from 10,000 paired resamples of 120 catalogue families, retaining their six conditions (seed 202609260540). Pointwise intervals and added/lost counts are archived for every contrast.

Contrast	High precision: Δ [adjusted CI]	Strict: Δ [adjusted CI]
N minus T	$+4.58\ [1.81, 7.92]$	$-0.69\ [-4.31, 2.36]$
E minus T	$+4.58\ [1.94, 7.92]$	$-5.14\ [-9.74, -0.97]$
N minus E	$+0.00\ [-0.69, 0.69]$	$+4.44\ [1.39, 8.06]$
$\lambda = 0.25$ minus N	$-3.61\ [-6.67, -1.23]$	$+0.69\ [-2.36, 4.31]$
$\lambda = 0.25$ minus T	$+0.97\ [0.14, 2.08]$	$+0.00\ [0.00, 0.00]$
$\lambda = 0.50$ minus N	$-2.64\ [-5.42, -0.56]$	$+0.83\ [-2.22, 4.44]$
$\lambda = 0.50$ minus T	$+1.94\ [0.56, 3.89]$	$+0.14\ [0.00, 0.69]$
$\lambda = 0.75$ minus N	$-1.11\ [-3.06, 0.28]$	$+0.83\ [-2.22, 4.44]$
$\lambda = 0.75$ minus T	$+3.47\ [1.39, 6.25]$	$+0.14\ [0.00, 0.69]$

X PREDICTION-RISK DISTRIBUTIONS BEYOND THE COMPLETE MEAN

Figure 31 plots the survival function of all returned risks on cohort B. Each curve retains all 4,608 conditions, including nonfinite predictions at infinite risk. The curves show the proportion of returns exceeding any specified error level; the positive far-right floor therefore retains the failures that make each complete mean infinite.

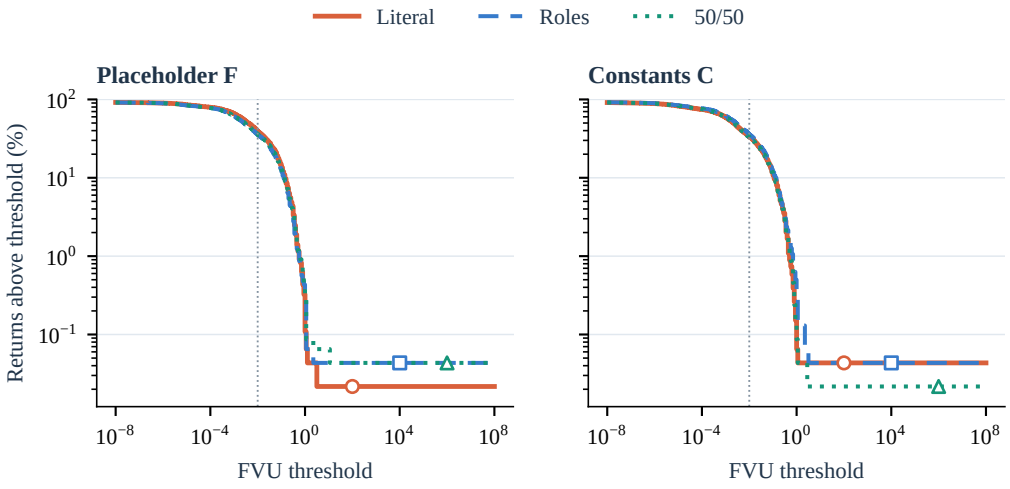


Figure 31: **Error distributions with failures retained.** Cohort B, three training seeds; log scales show both ordinary prediction errors and rare catastrophic returns. The vertical line marks FVU = .01. Line styles and markers distinguish configurations with coincident tail probabilities; labels give their nonfinite counts. The curves describe the original frozen returns across error thresholds.

Table 45: All-return risk and descriptive central-body summaries on cohort B. Quantiles use the inverse empirical CDF. The upper-5%-trimmed mean retains the lowest 4,377 risks and discards the largest 231 of 4,608, including nonfinite cases; it summarizes the lowest 95% of observed risks. The full mean and exceedance counts retain the tail. Exceedance counts include nonfinite risks.

Target / output	Full mean	95th	99th	Trimmed	$R > 1/10/10^3$	NF
Literals / F	∞	0.2445	0.5209	0.0260	5/1/1	1
Roles / F	∞	0.2085	0.5471	0.0217	15/2/2	2
50/50 / F	∞	0.2380	0.5999	0.0231	13/3/2	2
Literals / C	∞	0.2033	0.4559	0.0194	3/2/2	2
Roles / C	∞	0.2133	0.6440	0.0216	16/2/2	2
50/50 / C	∞	0.2035	0.5047	0.0188	4/1/1	1

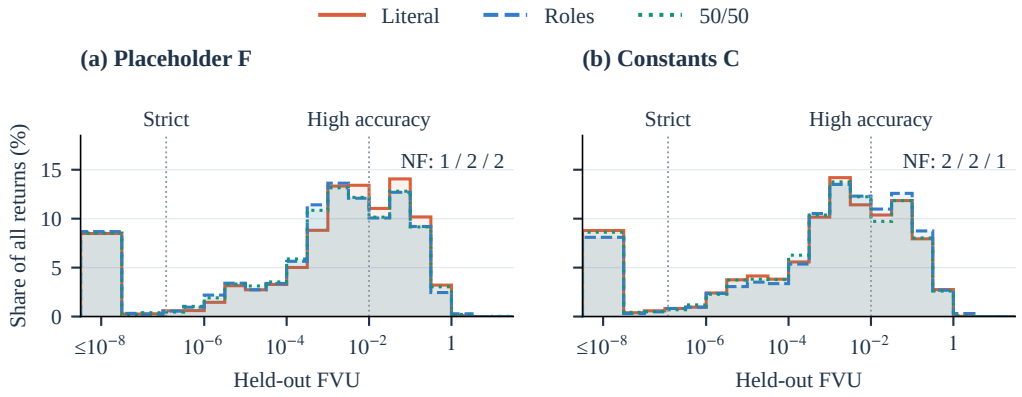


Figure 32: **The accuracy thresholds sit inside overlapping error distributions.** Cohort-B finite FVU, with each histogram bin divided by all 4,608 returns in that configuration. The leftmost bin pools FVU $\leq 10^{-8}$, including zeros; remaining bins have width 0.5 in $\log_{10}$ FVU. NF counts follow legend order and remain in all denominators; every complete mean is infinite. Vertical lines mark high accuracy and strict success. This is a descriptive view of the original three-seed confirmation, not another independent evaluation.

Y Bounded structural similarity on cohort B

We measure normalized tree-edit distance (NED) on all 27,648 frozen cohort-B returns: 256 families, six observation conditions, three training seeds, and six target–output cells. Following the original structural metric, symbolic canonicalization retains operator and variable labels and replaces every numeric node, including typed exponents, with `Const`. NED is the ordered-tree edit distance divided by the truth-tree node count and capped at one. Zero NED therefore records a match of the canonicalized numeric-free skeleton; coefficient recovery, algebraic equality, and domain validity are separate properties.

The symbolic evaluator uses a guarded one-second construction and distance pass. For 448 expressions requiring numeric-alias normalization, `np.e` and `np.pi` are represented by their binary64 values, with separate five-second construction and canonicalization/distance budgets. Canonicalization and the structural metric are shared across both paths. The complete evaluation provides 17,140 distances; 10,508 remain UNKNOWN because symbolic evaluation reaches a time or resource limit. UNKNOWN identifies returns whose symbolic evaluation exceeded its time or resource limit. For each cell, with $n = 4,608$ returns and u UNKNOWN outcomes, we report the full-denominator bounds

$$L = \frac{\sum_{i:\, d_i \text{ available}} d_i}{n}, \qquad U = L + \frac{u}{n}.$$

These are deterministic bounds obtained by assigning each UNKNOWN distance its full range $[0, 1]$. The six configurations have overlapping mean-NED bounds.

Table 46: Complete cohort-B structural diagnostic, $n = 4,608$ per row. F and C denote placeholder and constants outputs; Balanced uses the 50/50 literal/roles target-format mixture. NED$= 0$ counts only observed zero distances. UNKNOWN outcomes remain in the denominator and in the mean bounds; the final column partitions every UNKNOWN outcome. All native returns were nonempty. Lower NED indicates closer numeric-free tree structure.

Target	Output	NED$= 0$	Available	UNKNOWN	Mean NED bounds	Timeout / resource limit
Literal	F	482	2,852	1,756	$[0.3092, 0.6902]$	1,727 / 29
Literal	C	516	2,852	1,756	$[0.2894, 0.6704]$	1,736 / 20
Roles	F	512	2,858	1,750	$[0.2914, 0.6712]$	1,726 / 24
Roles	C	501	2,857	1,751	$[0.3038, 0.6838]$	1,723 / 28
Balanced	F	503	2,861	1,747	$[0.2919, 0.6710]$	1,718 / 29
Balanced	C	496	2,860	1,748	$[0.2932, 0.6726]$	1,729 / 19

Z Connections to recent hybrid recovery pipelines

MOSAIC-SR couples pretrained Transformer sketches to local symbolic repair and scale-aware constant optimization (Zheng et al., 2026). This division of work makes the proposal–refinement interface directly relevant: a sketch with free slots delegates numerical estimation, whereas an emitted coefficient can also initialize the optimizer. Our crossed design supplies a way to test these choices with the same backbone and observation stream. A transfer of the measured ranking would require crossing MOSAIC-SR's own targets and refinement routes; its reported recovery results and our coverage measurements use different protocols.

A-SR instead coordinates LLM proposal agents through evaluator feedback, role policies, and process memory, with optional distillation across runs (Zhao et al., 2026). Its agent roles denote search responsibilities, whereas our parameter roles denote numeric sites in expressions. Interface matching concerns the numerical content passed by such a proposal module to its fitter and can be studied alongside the orchestration policy.

PAN+SR uses nonparametric variable screening to reduce large input spaces before symbolic search (Ye & Li, 2025). This motivates a separate deployment question for fixed-width neural decoders: whether training-only feature screening turns unsupported high-dimensional inputs into useful returns. Our screening follow-up evaluates mutual-information, fixed, and random column subsets. It complements the target-format and output-interface interventions with a comparison of feature-selection rules.

AA Illustrative returned physical formulas

Verified returned laws	**Returned interior-pole failure**
I.39.11 : $\dfrac{x_2 x_3}{x_1 - 1}$	$\dfrac{0.001 x_1}{\cos^2(2.52911 x_3)}$
I.12.11 : $x_1(x_2 + x_3 x_4 \sin x_5)$	FVU 8.98×10^7; a pole near $x_3 = 10.55847$ lies inside the sampled domain $[10, 1000]$.
Algebraically equal on their natural domains. I.39.11 has a true endpoint pole.	

Figure 33: **What is returned matters beyond coverage.** Two verified laws and a role-scale seed-2 failure from the fixed physical panel. A true endpoint singularity and a spurious interior pole require different interpretation. These are illustrative saved returns, not additional experimental conditions.

AB DATA, TRAINING, AND CHECKPOINT CONTRACTS

Synthetic families. Each base expression uses one to four variables and 8–40 nodes. The six construction categories are additive expressions; products containing additive factors; rational expressions with denominator $1+b(x)^2$; sine/cosine combinations; $\exp(0.25\tanh(a(x)))+\log(1+b(x)^2)$; and products with a cosine factor. Subexpressions use addition, subtraction, multiplication, squaring, sine, cosine, and hyperbolic tangent. Atomic constants mix $\{-2,-1,-0.5,0.5,1,2\}$ with uniform draws on $[-3,3]$.

Canonical graph checks, including rational cancellation, remove coefficient-only and variable-renamed duplicates and exclude recorded catalogue and project exposures. The accepted 4,096 training and 1,024 development families are disjoint under this structural key. Its exclusions cover the recorded expressions; general algebraic equivalence and full upstream-pretraining overlap are outside that key's coverage.

For each family, 128 points are drawn uniformly from $[-1,1]^d$. The clean response must be finite with empirical standard deviation in $[10^{-7},10^{7}]$. A family noise level is drawn from $\{0,0.01,0.1\}$, and the same Gaussian perturbation is used in its four views. One view is identity; three independently sample input and output log-scales on $[-6,6]$. Native serialization must fit in 256 tokens and round-trip with prediction error below 10^{-8} in units of the base-response standard deviation. Rejected proposals enter neither split.

Optimization and target formats. T8-20M has approximately 23.7 million parameters. Continued models take 2,000 AdamW updates from shared pretrained weights, with weight decay 0.01, $\beta=(0.9,0.95)$, clipping at one, 100 warmup steps, and a cosine schedule from 10^{-5} toward 10^{-6}. One microbatch holds all four views of a family; four accumulated microbatches give sixteen view presentations per update, or 32,000 per arm. Identity continuation repeats the identity observations with the same dropout policy. Training uses BF16 autocast and the native numerical observation representation.

Loss is averaged over valid supervised tokens within each view, then over views. Padding and harness opening tokens are masked, and the native floating-target mask is retained. Roles replace only designated fittable sites; fixed exponents remain numerical. The training arrays contain 14,527 such fixed sites over four views in 2,555 of the 4,096 families. Target format changes prompts, sequence lengths, supervised-token counts, and realized training cost. The smooth group control uses the same views with

$$\mathcal{L}_{\text{group}} = (1-\rho)q^{-1}\sum_{g}\ell_g + \rho\tau\log\left(q^{-1}\sum_{g}e^{\ell_g/\tau}\right), \qquad \rho=\tau=0.5. \tag{5}$$

Its extra high-accuracy effects over ordinary augmentation have intervals spanning zero in the original three-seed confirmation.

Fixed checkpoints and earlier monitoring. Development CE is recorded every 200 steps on the first 128 development families. All main training comparisons use update 2,000. Earlier literal-only development and physical combinations used CE-best identity at step 1,000; their scale/group checkpoints were already the last checkpoint. The later step-matched identity row has 55 strict development successes, versus 52 for the earlier CE-selected checkpoint. Neither is selected by formula accuracy. The other 896 development families have also been inspected and are not an untouched confirmation set.

Execution records. The earlier local runs use an RTX 5090 Laptop GPU (24 GB nominal), an Intel Core Ultra 9 275HX, and 64 GB nominal host RAM. Flash-ANSR 0.18.0 uses PyTorch 2.11.0 and FP64 SciPy refinement. Native millibits are divided by 1,000 for reported lengths. Component costs include every needed source and exclude offline adaptation; source services ran under contention. Training, generation, refinement, and selection records remain separate. The companion arrays, checkpoint manifests, evaluator records, and table scripts identify each result without treating wall-clock differences as matched compute.

Table 47: Seed-0 checkpoint policies. Main comparisons use the fixed final step.

Training	Target	Final step	CE-best step
Identity	Literal	2000	1000
Scale	Literal	2000	2000
Group	Literal	2000	2000
Identity	Roles	2000	600
Scale	Roles	2000	2000
Group	Roles	2000	2000

AC COMPLETE 20M FAMILY AND OUTPUT-INTERFACE COMPARISONS

The 256 confirmation families were generated after development and exclude the 5,120 training/development families under the recorded key. They remain within the same grammar. Each model evaluates three noise levels and two coordinate views per family: 1,536 conditions with 128 visible and 512 clean audit points. The 18 continued models and two frozen references give 30,720 returns, with zero discrepancies under the independent operation-preserving evaluator. Later interface, coordinate, and capacity analyses reuse these exposed families.

Table 48: All original placeholder models, 1,536 conditions per row. NF counts nonfinite predictions; complete means retain them. Bits are mean output length. "Robust" in archived row labels denotes the smooth group objective.

Model / seed	$R^2 \geq .99$	Strict	Mean FVU	Median FVU	NF	Bits
frozen20	430	25	∞	0.05534	14	75.81
frozen120	660	51	∞	0.0179	11	84.05
Literals/identity / 0	829	172	1.01e+09	0.007171	0	57.39
Roles/identity / 0	847	180	∞	0.00595	2	54.78
Literals/identity / 1	799	163	0.1084	0.008348	0	57.33
Roles/identity / 1	851	177	0.08886	0.005935	0	55.79
Literals/identity / 2	816	163	0.1037	0.007587	0	58.46
Roles/identity / 2	826	176	∞	0.006337	2	55.77
Literals/scale / 0	944	167	∞	0.003786	1	63.55
Roles/scale / 0	993	188	0.04668	0.002639	0	63.81
Literals/scale / 1	949	166	0.04519	0.00346	0	63.63
Roles/scale / 1	996	169	0.04216	0.002672	0	65.42
Literals/scale / 2	949	162	0.04786	0.003325	0	64.85
Roles/scale / 2	1007	170	∞	0.002464	1	65.34
Literals/robust / 0	937	154	∞	0.003984	1	63.38
Roles/robust / 0	997	177	0.04218	0.002756	0	64.95
Literals/robust / 1	941	152	0.04715	0.003744	0	63.69
Roles/robust / 1	999	164	∞	0.002634	1	65.71
Literals/robust / 2	948	150	0.04771	0.003642	0	63.42
Roles/robust / 2	999	178	∞	0.002459	2	65.27

Paired intervals. We average noise conditions within each family, scale view, and training seed. Each of 10,000 bootstrap draws samples 256 families and three training seeds independently with replacement, crossing the indices and sharing them across arms. Each bootstrap replicate samples seeds globally across families. Intervals are pointwise for each contrast. The scale–role interaction is the role effect under scale training minus its effect under identity continuation.

Joint success and shell evaluation. For success in both recorded coordinate views, the 768 family–noise pairs per seed give literal identity counts 247 / 236 / 239, literal scale 376 / 381 / 385, and role scale 400 / 418 / 409. The literal scale increment is 18.23 points ($[14.37, 22.31]$); adding roles gives 3.69 ($[1.22, 6.25]$). Different search seeds make this a joint-success endpoint, not functional equivariance. Shell evaluation holds all 30,720 returned formulas fixed and uses

Table 49: All original $R^2 \geq .99$ contrasts, in percentage points. Positive differences favor the first intervention; original and changed coordinates each have 768 conditions per seed.

Contrast	Original (pp)	95% interval	Changed (pp)	95% interval
Scale effect, literals	-4.56	[-6.99, -2.13]	+21.83	[+17.66, +26.22]
Scale effect, roles	-2.99	[-5.34, -0.78]	+23.48	[+19.14, +27.99]
Role effect, identity	+1.74	[-0.43, +4.08]	+1.74	[-0.43, +4.17]
Role effect, scale	+3.30	[+1.30, +5.34]	+3.39	[+0.87, +5.99]
Group effect, literals	-1.56	[-3.52, +0.52]	+0.87	[-1.35, +2.78]
Group effect, roles	+0.13	[-1.39, +1.65]	-0.17	[-1.95, +1.78]
Scale–role interaction	+1.56	[-1.82, +4.82]	+1.65	[-1.78, +4.99]

$[-2, 2]^d \setminus [-1, 1]^d$ without refitting. Every model has infinite complete mean risk, with all nonfinite conditions retained below.

Table 50: All original models on the shell domain, 1,536 conditions per row. The returned formulas are unchanged from interpolation; all complete means are infinite. The last column reports the full-denominator median FVU, with nonfinite values retained as infinity.

Model	$R^2 \geq .99$	Strict	NF	Median FVU
frozen20	59	13	646	5.893
frozen120	93	27	614	2.998
literal_raw_ce	358	169	44	0.4917
roles_raw_ce	361	174	23	0.4033
literal_raw_ce_s1	334	160	32	0.4844
roles_raw_ce_s1	349	174	35	0.4241
literal_raw_ce_s2	341	157	35	0.4684
roles_raw_ce_s2	367	168	38	0.4366
literal_scale_erm	351	156	42	0.3614
roles_scale_erm	391	172	29	0.2411
literal_scale_erm_s1	351	154	26	0.3116
roles_scale_erm_s1	372	155	27	0.2601
literal_scale_erm_s2	342	146	36	0.3358
roles_scale_erm_s2	378	152	28	0.2289
literal_robust	343	142	38	0.3448
roles_robust	381	165	31	0.2426
literal_robust_s1	334	140	29	0.3327
roles_robust_s1	365	152	30	0.2577
literal_robust_s2	330	135	47	0.3588
roles_robust_s2	391	159	37	0.2264

AC.1 THREE-SEED CONSTANT EMISSION

The six scale checkpoints are unchanged: two target formats, three seeds, update 2,000. Constant emission uses the same task seeds and 64 proposals on all 1,536 conditions per model. Its 9,216 returns have zero independent-evaluator discrepancies. Numerically spelled refinement slots permit native predicted initialization; the attempted fits depend on each candidate and its visible fit criterion. This follow-up was fixed after the placeholder readout and changes no weights.

The adjusted high99 and high999 interaction intervals exclude zero; the strict interaction interval and constant simple-effect intervals include zero. Per-seed high-accuracy role effects are $+3.19, +3.06, +3.78$ points with placeholders and $-0.98, -2.21, -2.41$ with constants. Their interactions are $-4.17, -5.27, -6.18$. Original/changed strata remain in the complete machine table.

Table 51: All scale-trained interface results, 1,536 conditions per row. F/C denote placeholder/constant emission. NF remains in the full denominator and mean.

Target / seed	Format	$R^2 \geq .99$	Strict	Mean FVU	Median FVU	NF
Literal / 0	F	944	167	∞	0.003786	1
Literal / 1	F	949	166	0.04519	0.00346	0
Literal / 2	F	949	162	0.04786	0.003325	0
Roles / 0	F	993	188	0.04668	0.002639	0
Roles / 1	F	996	169	0.04216	0.002672	0
Roles / 2	F	1007	170	∞	0.002464	1
Literal / 0	C	1037	179	0.03538	0.001734	0
Literal / 1	C	1037	167	0.03658	0.001839	0
Literal / 2	C	1040	186	∞	0.001895	1
Roles / 0	C	1022	163	0.04093	0.002485	0
Roles / 1	C	1003	156	0.3052	0.002198	0
Roles / 2	C	1003	155	0.03937	0.00249	0

Table 52: Complete 20M interface effects (percentage points). The interaction is the role-minus-literal effect with constants minus that with placeholders. The fixed family contains these nine comparisons. Pointwise 95% and Bonferroni-adjusted intervals use the same 10,000 crossed family–seed draws.

Contrast	Endpoint	Difference	Pointwise 95%	Bonferroni (nine)
Roles − literal, F	$R^2 \geq .99$	+3.34	[+1.76, +4.90]	[+1.08, +5.53]
Roles − literal, C	$R^2 \geq .99$	-1.87	[-3.65, -0.04]	[-4.43, +0.88]
Interface interaction	$R^2 \geq .99$	-5.21	[-7.55, -2.80]	[-8.72, -1.56]
Roles − literal, F	$R^2 \geq .999$	+4.34	[+2.43, +6.32]	[+1.56, +7.21]
Roles − literal, C	$R^2 \geq .999$	-2.80	[-4.88, -0.72]	[-5.86, +0.13]
Interface interaction	$R^2 \geq .999$	-7.14	[-10.03, -4.30]	[-11.40, -2.99]
Roles − literal, F	Strict	+0.69	[-0.33, +1.76]	[-0.78, +2.33]
Roles − literal, C	Strict	-1.26	[-2.39, -0.20]	[-2.93, +0.33]
Interface interaction	Strict	-1.95	[-3.32, -0.46]	[-3.84, +0.39]

AC.2 Independent clean-point grids

We retain every eighth family in the original ordered cohort, fixing 32 already-exposed families without an outcome filter. The 2,304 existing returns cross four scale-trained configurations, three seeds, and six conditions per family. Each formula is evaluated on three independent 2,048-point uniform $[-1, 1]^d$ grids, transported with the saved coordinate scales. A family's grid is shared across methods, training seeds, and noise levels. No formula is regenerated, refitted, or reselected; this evaluates sensitivity to clean-point sampling on fixed structures.

Table 53: Fixed-formula grid stability. Each row retains 2,304 evaluations (576 per configuration). Interactions are percentage points. High99 flips count changed success labels relative to the original 512-point grid.

Grid	High99 interaction	High999 interaction	Strict interaction	High99 flips	NF / $R > 10$
1	-5.90	-9.20	-4.34	20	0 / 0
2	-6.08	-9.20	-4.34	23	0 / 0
3	-6.42	-9.38	-4.34	17	0 / 0

All three interaction endpoints are negative in every grid and training seed. There are no invalid ground-truth grids. This subset has no nonfinite return or $R > 10$ case; its per-configuration mean FVU ranges from 0.04020 to 0.04656. Nevertheless, 20/23/17 high99 labels change across the respective grids. These are repeated evaluations of the same 32 families, not three structural replications; the full-cohort nonfinite risks remain as reported above.

AD THREE-SEED 120M TRAINING AND EMISSION

The capacity comparison crosses identity/scale training, literal/role targets, and placeholder/constant emission over three training seeds, all at update 2,000. The 256 now-exposed families, six conditions per family, and 64-proposal contract match the 20M comparison. The complete suite contains 36,864 returns, including the 18,432 placeholder returns. Each pooled cell has 4,608 conditions; each original/changed stratum has 2,304. Two capacities replicate the interaction and scale-training comparisons.

Table 54: Complete 120M training–target–emission cells, 4,608 conditions each. F/C denote placeholder/constant emission. All failures and nonfinite risk remain.

Training	Target	Emission	High99	High999	Strict	NF	Median FVU	Mean FVU
Identity	Literal	F	2972	1976	817	0	0.002272	1218
Identity	Roles	F	2966	1994	843	3	0.00221	∞
Scale	Literal	F	3389	2239	852	0	0.001107	0.02898
Scale	Roles	F	3484	2419	873	1	0.0008456	∞
Identity	Literal	C	3063	2031	734	4	0.00196	∞
Identity	Roles	C	3086	2081	761	5	0.00176	∞
Scale	Literal	C	3755	2694	911	1	0.0005132	∞
Scale	Roles	C	3579	2427	795	1	0.000819	∞

Table 55: 120M constant emission by training seed, 1,536 conditions per row.

Training	Target	Seed	High99	High999	Strict	NF	Mean FVU
Identity	Literal	0	1009	669	240	2	∞
Identity	Literal	1	1035	684	244	1	∞
Identity	Literal	2	1019	678	250	1	∞
Identity	Roles	0	1012	681	262	1	∞
Identity	Roles	1	1047	700	247	2	∞
Identity	Roles	2	1027	700	252	2	∞
Scale	Literal	0	1251	892	300	0	0.01756
Scale	Literal	1	1255	902	298	1	∞
Scale	Literal	2	1249	900	313	0	0.01576
Scale	Roles	0	1192	793	265	1	∞
Scale	Roles	1	1186	799	266	0	0.02533
Scale	Roles	2	1201	835	264	0	0.02368

The scale-trained high99 interaction is -5.88 points ($[-8.44, -3.54]$), with per-seed interactions $-5.34, -7.36, -4.95$. High999 and strict interactions are -9.70 ($[-12.83, -6.64]$) and -2.97 ($[-4.99, -0.87]$). These are pointwise crossed family–seed intervals. Constant emission has three-seed high99 literal counts 1,251/1,255/1,249 and role counts 1,192/1,186/1,201. Literal-scale constants have one nonfinite condition and infinite complete mean, whereas literal-scale placeholders have none.

Training-regime interaction. The identity-trained high99 interaction is $+0.63$ points, compared with -5.88 after scale continuation. We compare these interactions directly within paired family–seed resamples, retaining the same six conditions and all eight training–target–emission cells. The resulting three-factor contrast $I_{\text{scale}} - I_{\text{identity}}$ is negative at both high-precision thresholds (Table 56). This exploratory cohort-A comparison identifies sensitivity to continuation exposure; cohort B provides the separately designed confirmation within the scale-trained setting.

Literal identity/scale totals at $R^2 \geq .999$ are 1,976/2,239, and role totals are 1,994/2,419. Strict totals are 817/852 for literal and 843/873 for roles. The three nonfinite role-identity returns and single role-scale nonfinite return occur in seed 1. All-condition means for those pooled arms are infinite. The separate physical follow-up is reported with its 72-condition denominator in Appendix AH.

Table 56: 120M scale-minus-identity difference in the supervision–interface interaction. Values are percentage points, with pointwise 95% intervals from 10,000 paired crossed family–seed bootstrap draws. All three endpoints use the same 36,864 frozen returns on exposed cohort A.

Endpoint	$I_{\text{scale}} - I_{\text{identity}}$	Pointwise 95% interval
$R^2 \geq .99$	-6.51	[-9.44, -3.56]
$R^2 \geq .999$	-10.39	[-13.43, -7.47]
$R \leq 2^{-23}$	-2.99	[-5.92, +0.11]

Table 57: Every 120M placeholder-emission training seed, 1,536 conditions per row. Means retain all nonfinite predictions; there are no removed rows.

Training / target	Seed	$R^2 \geq .99$	$\geq .999$	Strict	NF	Mean FVU	Median FVU
Identity / literal	0	977	656	272	0	0.0832	0.002308
Identity / literal	1	991	649	260	0	0.09079	0.002342
Identity / literal	2	1004	671	285	0	3653	0.002209
Scale / literal	0	1133	738	276	0	0.02849	0.00116
Scale / literal	1	1111	743	295	0	0.02938	0.001134
Scale / literal	2	1145	758	281	0	0.02908	0.001045
Identity / roles	0	987	661	283	0	0.08016	0.002261
Identity / roles	1	1002	673	283	3	∞	0.002064
Identity / roles	2	977	660	277	0	0.07305	0.002247
Scale / roles	0	1156	791	292	0	0.02863	0.0009077
Scale / roles	1	1155	817	285	1	∞	0.0008139
Scale / roles	2	1173	811	296	0	0.02944	0.0008137

AE FIXED-INITIALIZATION REPLAY AND ACTUAL WORK

Parents, data, and interventions. The replay uses 16 old development families at three noise levels and two coordinate views (96 conditions). It follows the earlier initialization diagnostics and is an exploratory mechanism comparison. All stable unique parents from each original 64-proposal cache are retained without an outcome filter. Constant- and placeholder-emission caches remain separate.

Each parameter-family/task pair shares a fixed random-start sequence R1–R9. R8 uses R1–R8; P8 uses the predicted free-coefficient vector p_0 and R1–R7; Q8 replaces p_0 by its signed nearest power-of-two magnitudes and uses the same seven random starts. R9 adds R9 to R8; P9 adds R8 to P8. Only FREE values change: typed exponents, named constants, parameter ties, and non-free roles stay fixed. The placeholder R8 arm uses its own full parent cache. Pooling both formats combines two 64-proposal caches, for a total of 128 proposals.

Every fit has a 200-expression-forward cap, including numerical-Jacobian probes, and a five-second wall limit. The common refinement has its recorded five-second post limit and 200-forward internal-fit cap. The best finite state reached before a cap remains eligible, with convergence tracked separately. Selection uses the same visible score; audit labels never choose a parent, initialization, or state. Native constant and placeholder returns retain their original restart, early-stop, and policies for refining typed constants. These references retain the native solver's historical fitting budgets and parameter policies.

Table 58: 120M placeholder-emission contrasts from the crossed family–seed bootstrap. Differences and intervals are percentage points.

Contrast	Coordinates	Endpoint	Difference (pp) [95% CI]
Literal scale−identity	all	$R^2 \geq .99$	+9.05 [+6.62, +11.46]
Literal scale−identity	original	$R^2 \geq .99$	-1.56 [-3.21, +0.00]
Literal scale−identity	changed	$R^2 \geq .99$	+19.66 [+15.28, +24.09]
Literal scale−identity	all	$R^2 \geq .999$	+5.71 [+3.65, +7.77]
Literal scale−identity	all	Strict	+0.76 [-0.85, +2.50]
Roles scale−identity	all	$R^2 \geq .99$	+11.24 [+8.46, +14.02]
Roles scale−identity	original	$R^2 \geq .99$	-1.39 [-3.47, +0.52]
Roles scale−identity	changed	$R^2 \geq .99$	+23.87 [+19.10, +28.65]
Roles scale−identity	all	$R^2 \geq .999$	+9.22 [+6.71, +11.68]
Roles scale−identity	all	Strict	+0.65 [-0.72, +1.97]
Roles−literal, scale	all	$R^2 \geq .99$	+2.06 [+0.46, +3.80]
Roles−literal, scale	all	$R^2 \geq .999$	+3.91 [+1.87, +6.03]
Roles−literal, scale	all	Strict	+0.46 [-1.11, +1.95]

Table 59: All final initialization returns, 96 conditions per row. NF is zero for every row. "Two-format union" is the explicit 128-proposal diagnostic.

Initialization / interface	Draws	$\geq .99$	$\geq .999$	Strict	Mean FVU	Max FVU	Mean bits
Random, 8 starts	64	59	29	12	0.05075	0.8491	62.1
Random, 9 starts	64	59	29	12	0.05072	0.8491	62.02
Predicted + 7 random	64	68	34	12	0.03889	1.002	67.8
Predicted + 8 random	64	68	34	12	0.04108	1.002	67.24
Coarse + 7 random	64	71	34	12	0.03795	1.064	68.28
Placeholders, 8 random	64	57	25	8	0.0432	1.032	60.99
Native constants	64	71	31	11	0.02625	0.3706	71.56
Native placeholders	64	58	28	8	0.04229	0.7915	63.12
Two-format union	128	66	34	12	0.03098	0.4342	64.62

Table 60: Paired family-bootstrap effects, percentage points, from 10,000 draws over the 16 development families. R denotes random starts, P predicted, Q coarse; the numeral gives total start positions. Cells report effect and 95% interval.

Contrast	$R^2 \geq .99$ (pp)	$R^2 \geq .999$ (pp)	Strict (pp)
P8 − R8	+9.38 [+3.13, +16.67]	+5.21 [+0.00, +11.46]	+0.00 [+0.00, +0.00]
Q8 − R8	+12.50 [+5.21, +21.88]	+5.21 [+0.00, +11.46]	+0.00 [+0.00, +0.00]
Q8 − P8	+3.13 [+0.00, +8.33]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
R9 − R8	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
P9 − P8	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
P8 − Native constants	-3.13 [-8.33, +0.00]	+3.13 [-5.21, +11.46]	+1.04 [-3.13, +6.25]
Q8 − Native constants	+0.00 [-6.25, +6.25]	+3.13 [-5.21, +11.46]	+1.04 [-3.13, +6.25]

Logical strategy cost. The strategies share start positions and per-fit limits; their actual work depends on convergence and refinement. The table counts all requested attempts, failed/capped fits, and refinement work. Logical costs include every component a complete strategy requires; cached physical execution is separately recorded. Service seconds sum generation and replay under concurrent execution. The native references lack historical forward counts. Fit failures and cap hits are counted separately from nonfinite final returns.

Table 61: Complete logical work for matched replay strategies. "No state" is a fit attempt with no accepted finite state; fit caps and forwards include failed attempts. Native references have no comparable historical forward ledger.

Initialization / interface	Fit attempts	Post attempts	Forwards	Fit caps	No state	Seconds
Random, 8 starts	39,448	28,575	5,372,456	7,671	2,096	9908.1
Random, 9 starts	44,379	28,411	5,799,965	8,669	2,348	10209.6
Predicted + 7 random	39,448	29,795	5,561,844	7,900	1,808	9933.9
Predicted + 8 random	44,379	29,799	5,986,278	8,872	2,100	10223.0
Coarse + 7 random	39,448	29,621	5,546,970	7,899	1,808	9920.4
Placeholders, 8 random	36,728	25,531	5,169,503	7,839	1,469	9354.9
Two-format union	76,176	54,106	10,541,959	15,510	3,565	19263.1

Continuous-parent, post-refinement, and native-return outcomes retain all 96 conditions. The coarse intervention retains the observed coverage benefit, while the native constant reference matches Q8 coverage at lower mean risk. These observations locate useful initialization information within this replay without identifying the entire native system's gain with one mechanism.

AE.1 Four output pools under common fitting limits

We cross both scale-trained targets and both emission prompts at seed 0 on 32 old conditions from six families (five contribute six conditions, one two). Each pool retains its first eight stable-unique parents in the original 64-proposal cache order, including unsupported positions. These are restricted parent pools, not new eight-proposal searches. R4 uses four shared Gaussian starts; P4 replaces the first with emitted p_0 for constant pools, retaining R4 when p_0 is unavailable. Parameter roles, the 200-forward/five-second fit caps, and the five-second refinement limit match the fixed-initialization replay.

Final R4 high-accuracy counts are 11/11 for literal F/C and 9/9 for role F/C. The interaction estimate is zero, with exploratory six-family interval $[-18.75, 18.75]$ percentage points. P4 changes constant counts to 12/9; the resulting interaction is -3.13 ($[-19.44, 12.50]$). The small, unequal family panel and restricted pools support a conditional pipeline comparison, with substantial uncertainty about the interaction.

Table 62: Four-pool replay. Top: final selections, all 32 conditions per row; NF and empty returns are zero. Bottom: full-rank jointly affine parents with at least one free parameter, comparing SVD to R4 continuous fits before the refinement. Both solvers have the printed high-accuracy/strict counts; the last column is their absolute mean-FVU difference.

Parent pool	Starts	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Mean FVU
Literal / F	R4	11	1	1	0.09205
Literal / C	R4	11	2	1	0.16265
Roles / F	R4	9	4	3	0.10201
Roles / C	R4	9	2	1	0.20626
Literal / C	P4	12	2	1	0.09797
Roles / C	P4	9	2	1	0.22354

Full-rank affine pool	Eligible parents	$R^2 \geq .99$	Strict	$\lvert\Delta\text{mean FVU}\rvert$
Literal / F	46	3	0	6.21×10^{-11}
Literal / C	45	4	0	6.33×10^{-12}
Roles / F	53	7	1	6.20×10^{-11}
Roles / C	34	2	0	3.56×10^{-11}

A conservative joint-affinity certificate identifies 178 full-rank, nonzero- parameter parents. SVD least squares and continuous R4 fitting have identical high-accuracy and strict counts in each pool, with mean-risk differences below 7×10^{-11}. This subset shows no simple linear-solver failure driving the coverage differences. All 1,024 parents remain in the stage records: generated, continuous,

converged-only, canonicalized, and final. The R4 continuous stages have 4/5/1/2 unavailable parents in literal F/C and role F/C; these remain recorded even though every task has a finite final selection.

AE.2 Empirical coefficient distributions as initializers

We compare Gaussian and empirical initializers on the complete Literal/C and Roles/C candidate pools from the same 16 development families, three noise levels, and two coordinate views used in the initialization study. One 20M training seed contributes 96 conditions. Each target retains all unique parents from its original 64 proposals: 5,442 literal and 5,445 role parents. The replay holds each pool, native parameter roles, fitting limits, postprocessing, and visible-data selection fixed while changing its starts.

Four starting policies. R uses eight independent Gaussian vectors with standard deviation 5. P replaces one with the candidate's emitted vector, retaining R when that vector is unavailable. E draws eight vectors with the same number of free coefficients from other parameterized expression families. The common bank pools emitted vectors from both targets for the same query; sampling first chooses an available target uniformly and then one of its eligible vectors uniformly, with replacement. Thus the bank includes both 64-proposal sources. Identical parameterized expressions in the two target pools receive identical E starts. Of 9,743 parents with free coefficients, 9,665 have same-query donors. The other 78 use donors from different ground-truth families under the same coordinate view. PE replaces E's first vector with the candidate's own prediction, retaining E when that prediction is unavailable. The 1,144 zero-free-coefficient parents use direct evaluation in every policy.

Each fitted start has a 200-expression-forward cap and a five-second wall cap. The original native refinement follows under its five-second limit. The final formula is selected on visible data, and all 768 selected returns are evaluated on the frozen 512-point clean grids. The bank and starts were fixed before fitting; final selections were fixed before this readout. This control measures the use of a common, query-conditioned coefficient distribution within the two saved candidate pools. Table 64 reports the actual fitting work separately from proposal generation.

Table 63: Empirical-initializer replay: all 96 conditions per row. The three success columns count $R^2 \geq .99$, $R^2 \geq .999$, and FVU $\leq 2^{-23}$. Every return is finite and nonempty; NF and fallback counts are zero in all eight cells. Mean, median, and 95th-percentile FVU retain every condition; the starting policies are defined above.

Target	Starts	.99	.999	Strict	Mean FVU	Median FVU	95th FVU
Literal	R	59	29	12	0.05075	0.00561	0.30460
Literal	P	68	34	12	0.03889	0.00237	0.23628
Literal	E	70	29	11	0.03505	0.00249	0.26415
Literal	PE	72	31	11	0.03029	0.00226	0.14924
Roles	R	64	26	9	0.05352	0.00409	0.34461
Roles	P	64	26	9	0.05354	0.00386	0.34461
Roles	E	72	30	10	0.03728	0.00278	0.28290
Roles	PE	72	30	10	0.03760	0.00285	0.28713

Transfer across expressions. With empirical starts, high-accuracy counts increase from 59 to 70 in the literal pool and from 64 to 72 in the role pool. The literal E–R difference is $+11.46$ percentage points with adjusted interval $[2.08, 22.92]$. Adding the candidate's own vector to these starts gives 72 in each pool; the literal PE–E difference is $+2.08$ points with adjusted interval $[0.00, 7.29]$, while the role success indicators are unchanged. The gains occur in high-accuracy coverage: E retains 29/30 returns at $R^2 \geq .999$ and 11/10 strict successes for literal/role training, compared with R's 29/26 and 12/9. These results show that numerical predictions can provide useful starting values across expressions. The pool-to-pool intervals in Table 65 quantify the remaining differences under the shared initializer distribution.

Paired intervals. The bootstrap resamples 16 ground-truth families in 10,000 paired draws, retaining all six conditions and eight arms of each sampled family. The seven primary comparisons

Table 64: Coordinate strata and actual fitting work in the empirical replay. Each original/changed cell has 48 conditions and shows $.99/.999/\mathrm{Strict}$ counts. Calls are mean expression forwards per condition, including continuous fitting and native refinement. Fit and post columns give mean optimizer attempts per condition in those stages. All policies request eight starts per parameterized parent; realized work depends on the parent and starting values.

Target	Starts	Original	Changed	Calls	Fit	Post
Literal	R	37/20/10	22/9/2	55,963.1	410.9	297.7
Literal	P	38/20/10	30/14/2	57,935.9	410.9	310.4
Literal	E	39/19/10	31/10/1	54,537.4	410.9	296.2
Literal	PE	39/19/10	33/12/1	56,627.2	410.9	311.3
Roles	R	38/16/8	26/10/1	68,429.0	401.0	349.9
Roles	P	38/16/8	26/10/1	68,326.4	401.0	347.8
Roles	E	38/16/8	34/14/2	65,048.4	401.0	354.4
Roles	PE	38/16/8	34/14/2	65,447.1	401.0	355.5

comprise P–R and PE–E within each target, literal–role differences under E and PE, and the difference between the two PE–E effects. Bonferroni-21 covers these comparisons at the three precision endpoints. E–R and PE–P within each target form four secondary comparisons with Bonferroni-12 adjustment. Table 65 includes both families of comparisons and all endpoints. The paired rows are exploratory evidence from exposed development families and one training seed.

Table 65: All planned empirical-initializer contrasts, in percentage points. "Literal–Roles, E" compares the targets under E; "Initializer interaction" is $(PE - E)_L - (PE - E)_R$. Primary rows use Bonferroni-21; secondary rows use Bonferroni-12, each across all three endpoints. Nominal and adjusted intervals use the same paired family draws. Zero-width intervals reflect identical observed success indicators in this panel.

Contrast	Endpoint	Difference	Nominal 95%	Adjusted 95%
Primary comparisons: adjustment over 21 tests				
Literal: P–R	.99	+9.38	$[+3.13, +17.71]$	$[+1.04, +21.98]$
Literal: PE–E	.99	+2.08	$[+0.00, +5.21]$	$[+0.00, +7.29]$
Roles: P–R	.99	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Roles: PE–E	.99	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Literal–Roles, E	.99	-2.08	$[-12.50, +8.33]$	$[-17.71, +13.64]$
Literal–Roles, PE	.99	+0.00	$[-8.33, +9.38]$	$[-12.50, +14.68]$
Initializer interaction	.99	+2.08	$[+0.00, +5.21]$	$[+0.00, +7.29]$
Literal: P–R	.999	+5.21	$[+0.00, +11.46]$	$[+0.00, +15.63]$
Literal: PE–E	.999	+2.08	$[+0.00, +6.25]$	$[+0.00, +10.42]$
Roles: P–R	.999	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Roles: PE–E	.999	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Literal–Roles, E	.999	-1.04	$[-11.46, +8.33]$	$[-18.75, +12.50]$
Literal–Roles, PE	.999	+1.04	$[-10.42, +11.46]$	$[-17.71, +15.63]$
Initializer interaction	.999	+2.08	$[+0.00, +6.25]$	$[+0.00, +10.42]$
Literal: P–R	Strict	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Literal: PE–E	Strict	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Roles: P–R	Strict	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Roles: PE–E	Strict	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Literal–Roles, E	Strict	+1.04	$[-2.08, +4.17]$	$[-4.17, +7.29]$
Literal–Roles, PE	Strict	+1.04	$[-2.08, +4.17]$	$[-4.17, +7.29]$
Initializer interaction	Strict	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
Secondary comparisons: adjustment over 12 tests				
Literal: E–R	.99	+11.46	$[+4.17, +19.79]$	$[+2.08, +22.92]$
Literal: PE–P	.99	+4.17	$[+0.00, +9.38]$	$[+0.00, +12.50]$
Roles: E–R	.99	+8.33	$[+0.00, +18.75]$	$[-3.13, +22.92]$
Roles: PE–P	.99	+8.33	$[+0.00, +18.75]$	$[-3.13, +22.92]$
Literal: E–R	.999	+0.00	$[-6.25, +6.25]$	$[-9.38, +10.42]$
Literal: PE–P	.999	-3.13	$[-8.33, +0.00]$	$[-10.42, +0.00]$
Roles: E–R	.999	+4.17	$[+0.00, +11.46]$	$[+0.00, +15.63]$
Roles: PE–P	.999	+4.17	$[+0.00, +11.46]$	$[+0.00, +15.63]$

Continued on next page

Table 65 (continued)

Contrast	Endpoint	Difference	Nominal 95%	Adjusted 95%
Literal: E–R	Strict	-1.04	$[-3.13, +0.00]$	$[-4.17, +0.00]$
Literal: PE–P	Strict	-1.04	$[-3.13, +0.00]$	$[-4.17, +0.00]$
Roles: E–R	Strict	+1.04	$[+0.00, +3.13]$	$[+0.00, +5.21]$
Roles: PE–P	Strict	+1.04	$[+0.00, +3.13]$	$[+0.00, +5.21]$

AF COORDINATES, REPEATED-SEARCH CONTROLS, AND DOMAINS

Catalogue protocol. Frozen120 placeholder generation covers all 120 SRSD catalogue families, two fixed observation draws, and three noise levels, giving 720 conditions. Each search uses 128 visible observations and 1,024 proposals. Raw and RMS searches have the same model, task data, native ranking, and refinement settings; the repeated search uses seed offset $1{,}000{,}003$. Paired policies pay 2,048 proposals. All visible choices and complete generation panels were sealed before the clean readout. The ledger retains 55 EXCLUDED families (330 conditions) and 65 PROVISIONAL families (390), pending domain and manual checks. Repeated searches reuse these same families.

Each selector chooses between two already completed winners. Native S1 compares their stored source-coordinate scores; error-only compares visible FVU with the same floor. Both require a finite eligible score, use the other source when only one is available, and prefer the first search on ties. Raw is first for mixed selection. The transported-S1 baseline retains its original eligibility. All output lengths price the full original-coordinate expression, including scale constants. A joint code for the view, transform metadata, and formula would require a separate stated coding convention. No invariant-MDL or Fisher-based parameter code is implemented by these selectors.

Table 66: Complete catalogue policies relevant to the main comparison, 720 conditions per row. Fail denotes selection failure. Single and paired rows have 1,024 and 2,048 proposals. Complete means retain nonfinite risk.

Pool / selection	Draws	$\geq .99$	$\geq .999$	Strict	Fail	NF	Mean FVU	Mean bits
Raw	1024	499	468	244	0	6	∞	64.72
RMS	1024	541	493	207	0	7	∞	460.5
Raw + raw	2048	517	489	260	0	5	∞	61.06
RMS + RMS	2048	547	508	217	0	5	∞	447.3
Raw + RMS, native S1	2048	585	541	284	0	5	∞	269.7
Raw + RMS, error	2048	585	542	252	0	6	∞	264.6
Raw + RMS, transported S1	2048	552	522	289	0	3	∞	104.2
Raw + RMS, native S1 + G2	2048	561	519	273	42	0	8.68×10^{35}	238.1
RMS + RMS, native S1 + G2	2048	533	492	206	54	0	1.43×10^{77}	397.8

Table 67: Catalogue exposure strata. Labels concern eligibility for new-family claims and remain provisional where recorded; all families remain evaluated.

Exposure stratum	Pool / selection	n	$\geq .99$	$\geq .999$	Strict	NF
Previously excluded (55)	Raw + raw	330	277	267	144	1
Previously excluded (55)	RMS + RMS	330	269	254	108	2
Previously excluded (55)	Raw + RMS, native S1	330	293	279	147	2
Previously excluded (55)	Raw + RMS, error	330	293	278	133	3
Previously excluded (55)	Raw + RMS, transported S1	330	287	273	150	1
Provisional (65)	Raw + raw	390	240	222	116	4
Provisional (65)	RMS + RMS	390	278	254	109	3
Provisional (65)	Raw + RMS, native S1	390	292	262	137	3
Provisional (65)	Raw + RMS, error	390	292	264	119	3
Provisional (65)	Raw + RMS, transported S1	390	265	249	139	2

Table 68: Catalogue paired effects (percentage points), resampling 120 families with all six conditions. Mixed means raw+RMS; native S1 is used unless another score is named. The 95% intervals are exploratory and per contrast.

Contrast	$R^2 \geq .99$ (pp)	$R^2 \geq .999$ (pp)	Strict (pp)
Mixed − two raw	+9.44 [+5.83, +13.33]	+7.22 [+3.89, +10.69]	+3.33 [+0.28, +6.39]
Mixed − two RMS	+5.28 [+2.50, +8.47]	+4.58 [+1.94, +7.50]	+9.31 [+5.56, +13.33]
Native − error, mixed	+0.00 [-0.42, +0.42]	-0.14 [-1.25, +0.97]	+4.44 [+2.22, +6.81]
Native − transported, mixed	+4.58 [+2.64, +6.67]	+2.64 [+1.11, +4.31]	-0.69 [-2.92, +1.39]
Two RMS − two raw	+4.17 [-0.97, +9.31]	+2.64 [-2.09, +7.50]	-5.97 [-11.39, -0.69]
G2 − no G2, mixed	-3.33 [-5.69, -1.39]	-3.06 [-5.00, -1.39]	-1.53 [-2.78, -0.56]
G2 − no G2, two RMS	-1.94 [-3.75, -0.42]	-2.22 [-3.89, -0.83]	-1.53 [-2.78, -0.56]

Admission and extreme finite risks. G2 acts on original-coordinate execution domains and admits only proved PASS expressions, including explicitly certified removable extensions. The admissible pool consists of these certified expressions. Its recorded budget is a 64-cell proof budget, depth eight, five seconds wall time, three CPU seconds, and 2 GiB per proof process. A failed selection returns the visible mean, stays in every denominator, and counts as a failure. Transported formula length is retained even when an admitted stable extension changes execution.

Native mixed selection has five nonfinite conditions and finite maximum FVU 4.23×10^{182}. Two raw searches have five and finite maximum 7.19×10^{265}. Native mixed G2 has no nonfinite final return but 42 selection failures, 561 high-accuracy successes, and mean FVU 8.68×10^{35}. Repeated RMS with G2 has 54 failures, 533 high-accuracy successes, and mean 1.43×10^{77}, with finite maximum approximately 1.03×10^{80}. Large finite prediction errors persist among admitted outputs. G2's UNKNOWN counts describe admission of individual source views; final-task failures are counted after selection.

Table 69: Catalogue cost sums over all 720 conditions. Logical totals charge every source required by a policy plus selection/admission. Service totals are recorded under contention; shared cached sources are counted once only in the separate physical execution ledger. These are not equal-time comparisons.

Pool / selection	Generation s	Refinement s	Gate s	Logical total s	Service total s
Raw	7714.1	52806.0	0.00	60520.2	61177.4
RMS	13311.3	86780.6	0.00	100091.9	101824.1
Raw + raw	22065.7	142394.6	0.00	164460.9	167539.5
RMS + RMS	20883.1	134500.9	0.00	155400.5	161705.5
Raw + RMS, native S1	21025.5	139586.6	0.00	160612.6	163002.0
Raw + RMS, error	21025.5	139586.6	0.00	160612.6	163002.0
Raw + RMS, transported S1	21025.5	139586.6	0.00	160612.6	163002.0
Raw + RMS, native S1 + G2	21025.5	139586.6	180.60	160793.2	163182.6
RMS + RMS, native S1 + G2	20883.1	134500.9	161.50	155562.0	161867.0

AF.1 Scales outside fine-tuning support

The synthetic control uses literal identity/scale checkpoints at 20M, all three training seeds, and the same 256 exposed families. Each family receives scale vectors with $|\log_{10} s|$ in $(6, 8]$ and $(8, 10]$, crossed with three noise levels. Each training setting has 4,608 conditions. Exponent directions are independent across coordinates; the two bands share directions but draw their magnitudes separately. Models share the original observations, transformed noise, scale vectors, and query seeds. These scales are outside fine-tuning support, whose observed components span $[-5.999823188, 5.999979415]$; upstream pretraining support is not established. Single searches use 64 proposals and paired searches 128.

The original single-search evaluation contains 18,432 returns across raw/RMS, both training modes, three seeds, and both bands. All four pooled complete means are infinite; 21 conditions contain 98 nonfinite points. Scale training raises raw-coordinate high-accuracy coverage in every seed and

Table 70: Complete coordinate policies on the two outside-scale bands, pooled over three seeds: 4,608 conditions per row. NF and infinite means are retained; bits price original-coordinate returns.

Training	Pool / selection	Draws	$\geq .99$	$\geq .999$	Strict	NF	Mean FVU	Median bits
Identity	Raw	64	843	324	57	14	∞	51.9
Identity	RMS	64	2095	847	112	5	∞	431.9
Identity	Raw + RMS, native S1	128	2198	924	117	4	∞	395.2
Identity	RMS + RMS	128	2395	1040	127	4	∞	432.1
Identity	Raw + RMS, error	128	2204	931	118	5	∞	397.9
Identity	RMS + RMS, error	128	2391	1054	127	5	∞	443.3
Identity	Raw + RMS, transported S1	128	1437	726	115	9	∞	57.6
Scale	Raw	64	1341	580	89	1	∞	58.6
Scale	RMS	64	2148	897	91	1	∞	437.7
Scale	Raw + RMS, native S1	128	2376	1109	120	1	∞	362.9
Scale	RMS + RMS	128	2445	1074	117	2	∞	436.2
Scale	Raw + RMS, error	128	2390	1124	123	1	∞	353.9
Scale	RMS + RMS, error	128	2450	1093	117	0	0.05979	446.8
Scale	Raw + RMS, transported S1	128	1698	874	120	1	∞	63.0

Table 71: Outside-scale paired contrasts, in percentage points. Intervals cross three training seeds with 256 families. The last two rows hold the same two-search pool type fixed while changing training.

Contrast	$R^2 \geq .99$ (pp)	$R^2 \geq .999$ (pp)	Strict (pp)
Identity: RMS pair − mixed	+4.28 [+2.67, +5.92]	+2.52 [+1.24, +3.84]	+0.22 [-0.07, +0.59]
Scale: RMS pair − mixed	+1.50 [-0.22, +3.19]	-0.76 [-2.47, +0.87]	-0.07 [-0.63, +0.54]
Scale − identity, mixed	+3.86 [+0.39, +7.70]	+4.01 [+0.98, +6.99]	+0.07 [-0.67, +0.80]
Scale − identity, RMS pair	+1.09 [-2.28, +4.62]	+0.74 [-2.32, +3.69]	-0.22 [-0.72, +0.26]

band. Its same-RMS increments have intervals spanning zero, and identity/RMS exceeds scale/raw in both bands. For raw strict success, the scale effects are $+0.911$ points ($[0.304, 1.606]$) in the first band and $+0.477$ ($[-0.087, 1.128]$) in the second. Same-RMS strict point effects are negative in both bands with intervals spanning zero. The later repeated-RMS comparison preserves this distinction between raw-coordinate training gains and gains over an effective normalization control.

Original-coordinate evaluation. The catalogue and outside-scale evaluations retain 9,936 cases. Native and independent execution agree on finite masks and numerical tolerance, including admitted extensions. Complete machine tables retain all remaining policies, per-band/exposure contrasts, noise summaries, and explicit undefined differences between infinite means.

AF.2 Earlier exposed-family discovery

Eight configurations on the already exposed 256-family cohort compare the same saved raw/RMS winners at 128 total proposals. For frozen120, transported S1 gives 720 high-accuracy successes, native S1 913, and error-only 918, versus 763 for two raw searches. Native S1 has 63 strict successes versus the two-raw 67; median original-coordinate length is 390.01 versus 83.79 bits. Its nine nonfinite conditions and infinite complete mean remain. The native-minus-two-raw high-accuracy effect is $+9.765625$ points ($[7.2265625, 12.4348958]$), resampling 256 families. This is an exposed discovery result, separate from the catalogue and outside-scale comparisons above.

Among 267 RMS-only high-accuracy cases, transported S1 rejects 207, native S1 four, and error-only six. Native S1 retains 98 of 108 raw-only successes; error-only retains 105. The pool contains useful complementarity, but changing the score can lengthen formulas and change risk. Across all eight configurations, error-only exceeds native S1 high-accuracy counts; it introduces a nonfinite return in literal-scale/F where native S1 has none. Full eight-configuration precision, risks, lengths, and source costs remain in the evidence snapshot.

AG STRUCTURAL READOUTS, VALUE SUPERVISION, AND NUMERICAL CONTEXT

Post-return structure. The 30,720 original placeholder returns have 1,692 proved-equal and 13,324 UNKNOWN equality outcomes. UNKNOWN is unresolved, not proved unequal. Normalized edit distance (NED) is computable for 19,070 returns, of which 3,594 have zero distance. NED collapses numerical nodes to `Const`, so zero distance records a match of the resulting operator-and-variable skeletons.

Table 72: Selected structural readouts, using all 1,536 conditions per model. NED bounds assign unavailable distances zero or one; available-subset means are not complete means. Full 20-model readouts remain in the evidence.

Model	Proved equal, seeds 0/1/2	NED available, seed 0	NED bounds, seed 0
Literal identity	113 / 104 / 98	948	[0.292, 0.675]
Literal scale	90 / 90 / 89	971	[0.299, 0.667]
Role scale	103 / 88 / 78	959	[0.277, 0.653]

FREE value-byte and boundary supervision. Whole-span and value-only controls share initialization, seed, all 4,096 training families, four views, sample order, optimizer, and 2,000 updates. They retain literal teacher inputs and the valid-token denominator. Whole-span deletion zeros the losses of eight value bytes and two boundary markers at each FREE site; value-only deletion preserves the boundary losses. Across the prepared training and development arrays (5,120 families, four views each), the mask marks 683,528 value-byte positions and retains supervision at 170,882 delimiter and 193,670 fixed-number positions. These are prepared-array counts, not cumulative training presentations.

Each deletion has its own uniform control scaling all valid weights to the same retained per-example mass. The gradients differ: value-only zeroing and its uniform control have mean pre-clipping norms 4.317 and 3.273, with all 2,000 steps above the clip threshold. Fixed final checkpoints use the same 64-proposal inference on 128 old development families (768 conditions). The value-only extension adds 3,072 returns, with zero independent-evaluator discrepancies.

Table 73: FREE-loss interventions. Each cell is high-accuracy / strict / NF, out of all 768 development conditions. Teacher digits remain unchanged.

Literal scale training loss	Placeholder	Constants
Full CE	454 / 64 / 0	500 / 70 / 1
Zero CE: 10-token FREE span	228 / 34 / 2	171 / 26 / 2
Uniform weight: span-matched	455 / 62 / 0	495 / 69 / 2
Zero CE: 8 FREE value bytes	457 / 59 / 0	494 / 69 / 0
Uniform weight: value-matched	454 / 59 / 0	487 / 65 / 0

Table 74: Complete FVU for the same loss controls. F/C denote placeholder/ constant emission; all 768 conditions remain in each mean and median.

Literal scale training loss	F mean	F median	C mean	C median
Full CE	0.05324	0.00442	∞	0.00205
Zero CE: 10-token FREE span	∞	0.06148	∞	0.13007
Uniform weight: span-matched	0.05250	0.00457	∞	0.00253
Zero CE: 8 FREE value bytes	0.05018	0.00427	0.04775	0.00269
Uniform weight: value-matched	0.05191	0.00443	0.04183	0.00259

Value-only zeroing preserves substantially more high-accuracy returns than whole-span zeroing. Its paired intervals include zero. Teacher digits, shared gradients, and pretrained numerical knowledge remain available in both controls, providing several routes through which numerical information can act.

Table 75: Value-only deletion minus full CE, percentage points, with paired 95% intervals resampling the 128 families and retaining all six conditions.

Emission	High accuracy: effect [95% interval]	Strict: effect [95% interval]
Placeholder	$+0.39\ [-1.95, 2.73]$	$-0.65\ [-2.21, 0.91]$
Constants	$-0.78\ [-3.26, 1.69]$	$-0.13\ [-1.95, 1.56]$

AG.1 Fixed typed numerical values

Two single-seed controls start from the same pretrained 20M model and use 2,000 role-scale updates with unchanged prompts, teachers, sample order, and inference. One zeros only the eight fixed value bytes, keeping both boundaries; the other uniformly matches retained valid-token weight. The fixed values are exponents $\{2, 3, 4, 5, 6, 8\}$. Placeholder emission uses 64 proposals on all 768 old development conditions. The 1,536 new returns join full-role CE and historical whole-span deletion references, which also retained teacher digits.

Table 76: Fixed typed losses, all 768 conditions per row. The first two thresholds are distinct from strict numerical success.

Role-scale training loss	$\geq .99$	$\geq .999$	Strict	NF	Mean FVU	Mean bits
Complete role CE	474	293	68	0	0.05157	62.72
Zero fixed value-byte CE	470	273	60	0	0.05238	62.92
Uniform matched token weight	480	275	61	0	0.05150	63.27
Zero fixed numeric-span CE	408	231	48	0	0.09318	60.05

Table 77: All planned fixed-value contrasts: effect and 95% family-paired interval in percentage points, 10,000 resamples of 128 families.

Contrast	Endpoint	Difference pp	95% interval
Value bytes − complete CE	$R^2 \geq .99$	-0.52	$[-2.34, 1.17]$
Value bytes − complete CE	$R^2 \geq .999$	-2.60	$[-4.69, -0.65]$
Value bytes − complete CE	Strict	-1.04	$[-2.34, 0.13]$
Uniform − complete CE	$R^2 \geq .99$	$+0.78$	$[-0.78, 2.34]$
Uniform − complete CE	$R^2 \geq .999$	-2.34	$[-4.95, 0.13]$
Uniform − complete CE	Strict	-0.91	$[-2.34, 0.39]$
Value bytes − whole span	$R^2 \geq .99$	$+8.07$	$[5.08, 11.20]$
Value bytes − whole span	$R^2 \geq .999$	$+5.47$	$[2.47, 8.59]$
Value bytes − whole span	Strict	$+1.56$	$[0.13, 2.99]$
Uniform − whole span	$R^2 \geq .99$	$+9.38$	$[6.25, 12.76]$
Uniform − whole span	$R^2 \geq .999$	$+5.73$	$[2.73, 8.85]$
Uniform − whole span	Strict	$+1.69$	$[0.39, 2.99]$
Value bytes − uniform	$R^2 \geq .99$	-1.30	$[-2.86, 0.13]$
Value bytes − uniform	$R^2 \geq .999$	-0.26	$[-2.86, 2.34]$
Value bytes − uniform	Strict	-0.13	$[-1.44, 1.30]$

Value deletion changes $R^2 \geq .99$ by -0.521 points ($[-2.344, 1.172]$), $R^2 \geq .999$ by -2.604 ($[-4.688, -0.651]$), and strict success by -1.042 ($[-2.344, 0.130]$). The value-only-minus-uniform intervals span zero at all three endpoints. The table retains contrasts with both full CE and whole-span deletion. All 3,072 returned formulas have finite predictions; none is an empty output or evaluation exception. Whole-span results remain a historical reference.

AG.2 Local gradients and teacher-forced context

The local gradient diagnostic uses 32 old families, four views, three checkpoints, evaluation mode, and no parameter updates. Median FREE-byte/rest gradient-norm ratios are 0.144 at initialization, 0.648 after literal scale training, and 5.142 after whole-span removal. Median cosines are 0.217,

0.0101, and 0.00191. These quantities describe gradients at the measured checkpoints. The rest bucket includes fixed typed numbers as well as structure.

The context diagnostic separately uses the initial20 and literal-scale models, 16 old training and 16 old development families, four views, and three input contexts: 192 groups and 768 sequences. One context rounds FREE log-base-two magnitudes while retaining sign and signed zero and clipping exponents to $[-1074, 1023]$. The other cyclically permutes values within split, view, and FREE role. Only eight FREE input bytes change; observations, original target labels, boundaries, typed spans, length, and padding remain fixed.

Targets strictly follow the first complete original FREE span. Pure structure contains expression operators, variables, and grammar tokens, excluding every numeric span, byte token, named constant, terminator, and EOS. Fixed typed value bytes form a separate bucket. Each context has 2,272 structural targets over 32 families and 904 typed targets over 21; the other 11 typed-family entries are unavailable, not zero. FREE targets and boundaries are excluded from both.

Table 78: Teacher-forced CE in nats per target. Original/coarse/permuted columns pool tokens; Δ_c, Δ_p equally weight family differences. Target sets are identical across contexts.

Model / target	Original	Coarse	Permuted	Δ_c	Δ_p
Initial20 / structure	3.520332	3.946472	3.628685	$+0.432012$	$+0.117869$
Initial20 / fixed values	0.358469	0.434196	0.359017	$+0.075025$	-0.000156
Literal scale / structure	0.490073	0.665655	0.620476	$+0.172678$	$+0.145778$
Literal scale / fixed values	0.024892	0.035094	0.028228	$+0.007658$	$+0.002866$

Literal-scale structural CE rises in 30/32 families under either change. On the 16 old development families, paired means are $+0.142015$ (16 increases) and $+0.080164$ (15 increases). Initial20 structure rises in 31/32 and 27/32. Typed values have mixed effects: initial20 permutation gives a family mean -0.000156, with 10 increases and 11 decreases. Its pooled-token CE changes in the opposite direction. An unchanged-input control reproduces native loss exactly. This measures fixed-checkpoint context sensitivity, whereas deleting target CE with teacher values retained measures a different axis.

AH PHYSICAL RECOVERY, SOURCE COMBINATIONS, AND RISK

Complete catalogue learning follow-up. The completed 120-family SRSD comparison has 20 configurations and 14,400 returns. Each configuration retains 720 conditions (two observation draws and three noise levels). Neural continuations use a single training seed, fixed update 2,000, placeholder emission, and 1,024 proposals. The same exposure ledger retains 55 EXCLUDED and 65 PROVISIONAL families. Raw and standardized inputs are fixed separate configurations; standardization includes centering and is not RMS conditioning. The following single-search learning comparison is separate from the frozen120 two-search policies in Appendix AF. Traditional methods retain their native preprocessing and budgets; the table compares the resulting system configurations.

Raw literal scale raises high-accuracy/strict counts from 461/243 to 504/277; roles raise them from 464/244 to 484/267. Group literal gives 506/275 and group roles 483/265. Thus the preferred averaging objective depends on the target format and endpoint in this training seed. The literal-scale mean remains 5.785×10^4 despite no nonfinite predictions; role-scale has one nonfinite return and infinite complete mean. These are point estimates without new paired intervals or additional training seeds on this panel.

Cohort and evaluation. The existing panel contains Feynman I.32.5, I.16.6, I.34.10, II.34.2, I.34.27, I.39.11, I.34.8, I.12.11, II.11.3, and component 0 of predator–prey, shear flow, and Van der Pol systems. ODE right-hand sides are sampled in boxes, not reconstructed trajectories. Physical laws use the source magnitude/sign distributions. Two observation draws and three noise levels form the original 72-condition source study. Matched adaptation uses draw one at 1% noise and draw two at 10%, giving 24 fixed conditions per model. Separate clean probe and audit sets each

Table 79: All 20 catalogue configurations, 720 conditions per row. Single neural searches use 1,024 proposals. Complete means retain all failures; strict numerical success uses the prediction-risk threshold; algebraic checks are reported on their specified subsets.

System / coordinates	$R^2 \geq .99$	$R^2 \geq .999$	Strict	Median FVU	Mean FVU	NF
Frozen 120M / raw	499	468	244	2.511×10^{-5}	∞	6
Frozen 120M / standardized	290	196	42	0.0468	∞	9
Frozen 20M / raw	456	409	211	1.959×10^{-4}	∞	2
Frozen 20M / standardized	236	152	30	0.1439	∞	8
Literal identity / raw	461	428	243	1.064×10^{-4}	∞	4
Literal identity / standardized	244	149	32	0.0981	∞	3
Literal scale / raw	504	467	277	1.346×10^{-5}	5.785×10^{4}	0
Literal scale / standardized	261	162	32	0.0602	∞	4
Roles identity / raw	464	433	244	8.755×10^{-5}	∞	1
Roles identity / standardized	265	164	34	0.0522	∞	7
Roles scale / raw	484	448	267	3.387×10^{-5}	∞	1
Roles scale / standardized	291	173	34	0.0350	∞	6
Literal group / raw	506	470	275	1.002×10^{-5}	∞	1
Literal group / standardized	270	167	28	0.0577	∞	6
Roles group / raw	483	448	265	4.591×10^{-5}	∞	3
Roles group / standardized	290	177	36	0.0352	∞	6
PySR	438	341	128	0.0016	∞	17
Operon	434	312	137	0.0035	∞	13
TIR	285	259	83	0.0884	2.025×10^{47}	0
GenSR / standardized	341	229	47	0.0130	9.930×10^{37}	0

have 2,048 points. Three training seeds give 72 repeated conditions per training arm, not 72 independently sampled physical laws.

I.39.11 has a response pole at $x_1 = 1$, an endpoint of its $[1, 2]$ interval. II.11.3's continuous sampling distribution crosses a response pole and need not have a finite second moment. Their FVU is the ratio computed on the finite evaluation sample, whose second moments are well defined.

Table 80: All 20M matched physical repeats: 24 conditions per model, update 2,000, 1,024 proposals, and neural return alone. Observations stay fixed across training seeds; means retain every return.

	Strict /24			Mean FVU		
Training and supervision	Seed 0	Seed 1	Seed 2	Seed 0	Seed 1	Seed 2
Identity, Literals	6	5	5	0.1909	0.2152	0.4183
Scale, Literals	8	8	7	28.73	0.1258	0.2519
Group, Literals	8	8	9	0.1283	3.515e+04	0.1307
Identity, Roles	6	5	5	0.1595	0.1558	0.5441
Scale, Roles	8	9	7	0.5508	0.1447	3.743e+06
Group, Roles	7	9	7	3.758	0.181	996.7

Algebraic identities and direct counterexamples. Independent prefix parsing and symbolic subtraction verify the returned laws in Figure 33. Identity continuation already recovers I.12.11; the extra exact law from scale training is I.39.11. I.12.11 uses five variables, beyond the adaptation grammar's four, and I.39.11 has a true pole unlike its bounded-denominator templates. Pretrained weights already support these operators and variable counts. Other nearly zero FVU values, including approximate physical constants and a saturated hyperbolic tangent, retain their numerical interpretation.

Role-scale seed 2 returns $0.001x_1/\cos^2(2.52911x_3)$ on one condition, with FVU 8.98×10^7 and complete mean 3.74×10^6. A denominator zero occurs near $x_3 = 10.55847$ inside $[10, 1000]$. Earlier literal scale/0 and group/1 means are 28.7289 and 35147.8002. Across the six earlier literal

Table 81: 120M physical follow-up pooled over three training seeds, 72 conditions per row. G denotes the historical finite-point guard. "Library" denotes the single archived finite-dictionary branch. "Two neural" adds the same frozen120 return, paying both 1,024-draw searches. The cached library is shared across these source combinations.

Training / target	Sources	$\geq .99$	$\geq .999$	Strict	Mean FVU	$R > 10$	NF
Identity / literal	Native	58	48	23	0.7398	2	0
Identity / literal	Native + G	58	48	23	0.7398	2	0
Identity / literal	Neural + library + G	59	55	25	0.2067	0	0
Identity / literal	Two neural + G	58	55	30	0.4974	1	0
Identity / literal	Two neural + library + G	59	58	31	0.2065	0	0
Scale / literal	Native	62	58	25	5.693	2	0
Scale / literal	Native + G	62	58	25	5.693	2	0
Scale / literal	Neural + library + G	62	59	25	1.445	1	0
Scale / literal	Two neural + G	62	62	31	1.467	1	0
Scale / literal	Two neural + library + G	62	62	31	1.405	1	0
Identity / roles	Native	53	49	23	4.54×10^{10}	1	0
Identity / roles	Native + G	53	49	23	4.54×10^{10}	1	0
Identity / roles	Neural + library + G	58	54	23	4.54×10^{10}	1	0
Identity / roles	Two neural + G	58	57	29	4.54×10^{10}	1	0
Identity / roles	Two neural + library + G	59	58	29	4.54×10^{10}	1	0
Scale / roles	Native	61	58	28	41.82	1	0
Scale / roles	Native + G	61	58	28	41.82	1	0
Scale / roles	Neural + library + G	61	58	28	41.8	1	0
Scale / roles	Two neural + G	61	61	34	0.3974	0	0
Scale / roles	Two neural + library + G	61	61	34	0.3036	0	0

scale/group models, three of 144 returned formulas have FVU above 10, two above 100, and one above 1,000. None is nonfinite on the sampled audit points.

The scale/0 II.11.3 denominator is $q_s = (ax_4 + 305700000 - x_5/3)^2 - x_3x_5/x_4$, with numerator $-0.024x_5$. At $a = 0.2$, $x_4 = 2{\times}10^9$, and $x_3 = 10^{-27}$, set $c = 3(ax_4{+}305700000)$. Then $q_s(c) < 0$ and $q_s(c \pm 1) > 0$, proving two interior zeros with nonzero numerator by continuity. The signs hold for the stored binary64 constant and exact $1/5$; sampled FVU is 687.42. For the group/1 return $x_4 \cos(x_1/5{-}x_3/4)/(7q_g)$ with $q_g = 6x_2{-}2(x_4/3{-}11x_5/32)^2$, legal $x_2 = 10^{-8}$ and $x_5 = 2{\times}10^9$ give poles at $x_4 = 33x_5/32 \pm 3\sqrt{3x_2}$; $x_1 = 10^{-10}, x_3 = 10^{-27}$ keep its numerator nonzero. FVU is 843534.15. In contrast, the group/1 I.34.8 return $6x_1(30x_2 - 242000000)^2/((300x_2 + 2x_4)^2 + 4)$ is bounded on its box yet has FVU 12.05.

AH.1 Common-pool domain replay

For each of the 20 models, both arms receive the same four sources: its neural return, the frozen120 return, the earlier library return, and the visible mean. They share visible S1 and the node-count tie break. G2 additionally admits only proved PASS executions, including certified removable extensions. All 480 comparisons reuse saved formulas without search or refitting. UNKNOWN denotes a return whose domain admissibility is unresolved.

Every model loses strict successes. Role-scale seed 2's mean changes from 3.74×10^6 to 0.511 and strict count from 11 to 7 in this combined pool. G2 also rejects the exact I.39.11 target at its true endpoint pole, while the bounded group/1 error survives. Uniform boundedness and exact recovery require compatible declared domains. Prediction accuracy is evaluated separately after domain admission.

Table 82: Every model under the common-pool G2 replay, 24 conditions per row. Before/after uses the same four-source pool, not the neural-only return. Gain/loss counts refer to strict numerical success.

Model in pool	Strict before	After	Mean before	After	Gain	Loss
frozen20	9	7	0.1256	0.2823	0	2
frozen120	7	5	0.2136	0.3693	0	2
literal_raw_ce	10	8	0.1462	0.275	0	2
roles_raw_ce	10	8	0.1465	0.2426	0	2
literal_raw_ce_s1	10	8	0.1471	0.3285	0	2
roles_raw_ce_s1	9	7	0.1463	0.2751	0	2
literal_raw_ce_s2	10	8	0.1462	0.3403	0	2
roles_raw_ce_s2	9	7	0.2728	0.3438	0	2
literal_scale_erm	12	8	28.73	0.297	0	4
roles_scale_erm	11	7	0.5713	0.6408	0	4
literal_scale_erm_s1	12	8	0.1458	0.2905	0	4
roles_scale_erm_s1	12	8	0.1623	0.2973	0	4
literal_scale_erm_s2	11	7	0.2727	0.3786	0	4
roles_scale_erm_s2	11	7	3.743e+06	0.5107	0	4
literal_robust	11	7	0.1488	0.3022	0	4
roles_robust	11	7	3.778	3.848	0	4
literal_robust_s1	11	7	3.515e+04	0.676	0	4
roles_robust_s1	11	7	0.2012	0.2777	0	4
literal_robust_s2	12	8	0.1515	0.2951	0	4
roles_robust_s2	11	7	996.8	15.55	0	4

AI PUBLIC AND GROUPED PREDICTIVE EVALUATION

The public short-budget slice uses 24 SRBench-2025 datasets and three official splits, with all 72 test splits retained. Native output coverage is reported separately from complete-system prediction: Flash accepts at most 17 inputs and GenSR ten; unsupported inputs use a labeled training-mean fallback. Mean/std standardization is training-only and is distinct from the RMS conditioning used in coordinate experiments. TIR keeps its author's nonfinite-to-zero prediction semantics. Neural systems use 1,024 proposals; other systems keep their recorded native budgets. These observed-target results are not symbolic-recovery counts or the official 30-repeat, 3,600-second competition protocol.

Table 83: All 24 public configurations, with 72 splits per row and dataset-equal means. Native coverage excludes fallbacks; NF stays in complete means. Neural continuations use one fixed checkpoint each. All coordinate choices are retained; the 24 systems give 1,728 evaluated returns.

System / coordinates	$R^2 \geq .99$	Median FVU	Mean FVU	Native	NF
Frozen 120M / raw	17/72	0.3672	0.5124	69/72	0
Frozen 120M / standardized	1/72	0.4172	6.6245	69/72	0
Frozen 20M / raw	17/72	0.4436	0.5525	69/72	0
Frozen 20M / standardized	1/72	0.4826	1.1082	69/72	0
Literal identity / raw	13/72	0.3556	0.6432	69/72	0
Literal identity / standardized	3/72	0.3736	0.8595	69/72	0
Literal scale / raw	15/72	0.3145	0.7931	69/72	0
Literal scale / standardized	3/72	0.3866	0.8146	69/72	0
Roles identity / raw	13/72	0.3209	0.8688	69/72	0
Roles identity / standardized	2/72	0.3738	0.8197	69/72	0
Roles scale / raw	16/72	0.2364	0.3588	69/72	0
Roles scale / standardized	4/72	0.3614	0.8503	69/72	0
Literal group / raw	17/72	0.3234	0.3776	69/72	0
Literal group / standardized	2/72	0.4040	1.5157	69/72	0

Continued on next page

Table 83 (continued)

System / coordinates	$R^2 \geq .99$	Median FVU	Mean FVU	Native	NF
Roles group / raw	16/72	0.3085	0.3763	69/72	0
Roles group / standardized	3/72	0.3733	0.8738	69/72	0
PySR	10/72	0.2670	2.6482	72/72	0
Operon	12/72	0.2262	25.6476	72/72	0
ParFam feasible	8/72	0.4282	∞	58/72	1
Ridge	0/72	0.4844	1.0562	72/72	0
Random forest	5/72	0.2549	0.4185	72/72	0
Histogram gradient boosting	0/72	1.0016	1.1979	72/72	0
TIR	7/72	0.2268	0.4633	72/72	0
GenSR / standardized	8/72	0.2840	1722.34	66/72	0

AI.1 Group-disjoint Concrete and Airfoil

Concrete has 1,030 rows in 427 formulation groups, keeping age measurements together. Airfoil has 1,503 rows in 106 condition groups, keeping frequency scans together. Three outer group folds use training-only internal selection. Neural methods receive all outer training rows and 1,024 proposals; adapted models use seed-0 update-2,000 weights. Raw and standardized configurations are fixed separately. Pooled RMSE aggregates held-out squared errors before the square root; fold SD is descriptive variation. One original ParFam return is unavailable, and its labeled training-mean fallback remains in the errors.

Table 84: All 23 configurations on both grouped datasets, three outer folds each. Native coverage and complete-system errors are reported separately.

	Concrete		Airfoil		Native returns
Method	RMSE	Fold SD	RMSE	Fold SD	Concrete, Airfoil
Frozen120, raw	8.562	0.260	4.602	0.299	3/3, 3/3
Frozen120, std	12.778	1.774	5.228	0.391	3/3, 3/3
Frozen20, raw	9.480	1.348	4.810	0.408	3/3, 3/3
Frozen20, std	11.833	0.536	5.428	0.172	3/3, 3/3
Identity/literal, raw	9.154	0.352	4.422	0.532	3/3, 3/3
Identity/literal, std	11.228	1.646	5.040	0.349	3/3, 3/3
Scale/literal, raw	8.560	0.084	4.444	0.446	3/3, 3/3
Scale/literal, std	10.860	0.153	5.075	0.274	3/3, 3/3
Identity/roles, raw	9.145	0.125	4.451	0.460	3/3, 3/3
Identity/roles, std	11.367	1.338	5.112	0.650	3/3, 3/3
Scale/roles, raw	8.120	1.087	4.831	0.549	3/3, 3/3
Scale/roles, std	11.037	1.577	4.967	0.122	3/3, 3/3
Group/literal, raw	8.540	0.679	4.598	0.351	3/3, 3/3
Group/literal, std	11.080	0.699	5.109	0.647	3/3, 3/3
Group/roles, raw	8.221	0.275	4.437	0.507	3/3, 3/3
Group/roles, std	10.816	0.480	5.015	0.230	3/3, 3/3
PySR	7.661	0.341	5.088	0.285	3/3, 3/3
Operon	6.985	0.593	3.248	0.234	3/3, 3/3
ParFam, original	58.973	39.525	8.342	4.648	2/3, 3/3
ParFam, feasible configuration	64.768	22.813	9.513	3.893	3/3, 3/3
Ridge	10.791	0.605	4.925	0.176	3/3, 3/3
Random forest	6.350	0.814	2.927	0.533	3/3, 3/3
Histogram gradient boosting	6.134	0.565	2.973	0.597	3/3, 3/3

An earlier fixed neural/library combination has Vineyard pooled RMSE 3.6116, identical to neural alone, versus ridge 3.1400 and random forest 3.4107. On Pollution it has 63.5086 versus neural 54.3754, ridge 43.7825, and random forest 42.1523. These earlier source-combination results concern different systems and are not assigned to the adapted checkpoints.

AJ EARLIER CONTROLS AND ARCHIVED FULL COMPARISONS

Development controls. The earlier seed-0 comparison uses 128 exposed families and 768 conditions. Role-minus-literal scale has 56 gains and 36 losses, with family-bootstrap interval $[-2, 43]$ net conditions. The later three-seed confirmation supplies main inference. Literal ranges $[-2, 2]$ and $[-10, 10]$, and two unique views repeated, give 443, 457, and 447 high-accuracy successes versus the main four-view scheme's 454; their paired intervals span zero. The historical fixed-exponent-span deletion gives 408 and removes boundaries as well as values.

On this development cohort, transported raw/RMS selection gives frozen120 341 high-accuracy and 17 strict returns, versus 359/19 for two raw searches at the same 128 proposals. Later exposed-family selector outcomes appear in Appendix AF.2. Binary64 RMS encoding is not bitwise invariant: 1,152 recorded normalized pairs differ by at most 1.78×10^{-15} in mantissas. No replacement pre-encoder is assumed; native padded variable positions remain in transported expressions.

Saved-pool score sensitivity. Four seed-0 checkpoints contribute all retained candidates on 768 development conditions each. Error floors $\{2^{-52}, 10^{-12}, 2^{-23}\}$ cross length multipliers $\{0, 0.5, 1, 2\}$. Native settings reproduce all 3,072 winners. Changing floors at fixed weight leaves high-accuracy/strict counts unchanged, although some formulas change. Zero length weight gives 6/4/10/1 nonfinite returns across the four models; native weight gives none. Final-pool reranking selects among the retained final candidates.

Archived library and source controls. The finite dictionary includes degree-at-most-four monomials and products of quadratic monomials with one unary factor, capped at 5,000 columns. Target views are $y, y^2, y^3, \exp(y), \log(y), 1/y$ where admissible, with inverse links. Beam width/branching/depth are eight; coefficients use scaled FP64 least squares. Extra fitting allows at most eight free coefficients and 100 nodes, followed by four refinement shortlists per stream. The historical G guard checks 2,048 scrambled Sobol points and can miss a pole. FFX supplies its author frontier (McConaghy, 2011); PySR, Operon, and ParFam are separate native-budget systems, not branches of this dictionary.

On the original 72-condition physical source study, neural/library/combined strict counts are 32/18/37 and means 0.268855/0.4296/0.181142. Two added noisy II.34.2 returns equal $x_1x_2x_3/2$. At $R^2 \geq .999$ and .99, combined/neural counts are 59/58 and 60/59. Eight lower and seven higher continuous risks make the mean effect heterogeneous. The FFX combination's mean gain concentrates in one condition and fails to replicate in its fresh whole-system repeat. Full-menu versus identity-only dictionary reruns give 37/36 strict returns; removing inverse-origin states loses one, while other single-view removals change no output.

The old coordinate study gives 37 neural strict successes for two raw searches and 39 for raw/RMS at 8,704 draws; adding the shared library gives 40/41. The second raw source alone has mean FVU 1.79×10^{15} and unfiltered RMS 6.03×10^{92}. Twelve same-parent refinement comparisons each retain five strict successes with mean lengths 74.5247 versus 54.0908 bits. These historical budgets and failure tails stay separate from the new catalogue.

Program selection and complete archive. An exposed sixteen-condition comparison reruns five programs in four inner folds (96 training, 32 held-out observations), then regenerates the selected program on all 128 observations. Minimum OOF risk and the one-SE rule choose the same formulas as fixed mixed beam, with unchanged strict count; mean component cost rises from 107.2 to 1,249.6 seconds. Author TIR's 90/90 inner split remains within outer training data. The archive retains every historical per-arm row, score setting, development output-format result, shell row, and source-cost ledger. Current main comparisons and their direct contrary controls are printed in full; manifests and scripts preserve the historical detail.

AK COORDINATE AND PHYSICAL-RECOVERY STUDY

The coordinate catalogue covers all 120 SRSD families, two sampling draws, and three noise levels: 720 conditions. Its exposure ledger marks 55 families EXCLUDED from new-family claims and 65

PROVISIONAL pending domain/manual checks. The scale-range control reuses the 256 synthetic families at $|\log_{10} s| \in (6, 8]$ and $(8, 10]$, outside fine-tuning support.

RMS conditioning uses visible-data column scales without centering. For $z_i = x_i/s_i$ and $w = y/t$, transport is

$$\mathcal{T}_s[h](x) = t\, h(x_1/s_1, \ldots, x_d/s_d). \tag{6}$$

FVU is invariant for the same function under exact-arithmetic transport; its description length and discrete refinement depend on the coordinates. We compare saved raw/RMS returns by native S1, visible error, or transported S1. Repeated-coordinate searches provide the matched-proposal control. All lengths price the original-coordinate formula; complete rules and the score-transport relation appear in Appendix E.

AK.1 Coordinate choices and physical recovery

Training and searching across scales. Paired scale training improves raw-coordinate placeholder coverage at both capacities in every seed, mainly on changed coordinates, with modest original-scale losses. Strict and proved-equal recovery do not improve consistently. Literal-scale 120M has finite mean FVU 0.02898; role-scale has one nonfinite return and an infinite complete mean. A matched non-scale diversity control still favors scale training on changed coordinates (Appendices AB, AD, and H).

At 2,048 proposals, raw+RMS native-score selection returns 585/720 high-accuracy formulas, versus 517/547 for repeated raw/RMS searches. Strict counts are 284 versus 260/217: paired gains of $+3.33$ points ($[0.28, 6.39]$) and $+9.31$ ($[5.56, 13.33]$). Error-only and transported-score rules return 252 and 289 strict successes. Outside fine-tuning support, repeated RMS outperforms mixed coordinates under identity training; the scale-trained comparison has an interval spanning zero. Full coordinate, length, work, and scale-range comparisons appear in Appendices AF and AF.1.

Returned physical formulas. On 24 fixed physical conditions, scale-trained models recover I.39.11, $x_2x_3/(x_1-1)$, and I.12.11, $x_1(x_2+x_3x_4\sin x_5)$, at both noise settings in every seed; identity training already recovers the latter. Independent parsing and subtraction verify equality. A saved counterexample, $0.001x_1/\cos^2(2.52911x_3)$, introduces an interior pole and has FVU 8.98×10^7. At 120M, literal coverage rises from 58 to 62/72 while mean FVU rises from 0.7398 to 5.6931. Appendices AA and AH retain formulas, repeats, and domain-admission checks.

Coordinate and dataset choices matter. Seed-0 standardization reduces role-scale to 4/72 high-accuracy returns. On grouped Concrete/Airfoil folds, its RMSE is 8.120/4.831, versus frozen120's 8.562/4.602 and random forest's 6.350/2.927 (Appendix AI.1).

AL Additional related-work context

Learning and fitting expressions. Flash-ANSR provides amortized generation, native parameter roles, and numerical refinement (Saegert & Köthe, 2026). NeSymReS predicts skeletons and fits coefficients (Biggio et al., 2021); E2E and SymFormer predict constants to initialize refinement (Kamienny et al., 2022; Vastl et al., 2024). MMSR jointly optimizes symbolic cross-entropy, constant regression, and modal alignment (Li et al., 2025). Our comparison crosses target format with output emission inside one pretrained system, then isolates initialization under fixed parent pools and fitting limits. Unified deep symbolic regression and RAG-SR combine learned proposals with other search procedures (Landajuela et al., 2022; Zhang et al., 2025). RAG-SR pairs an output with both signs and obtains coefficients by linear regression; our training views rescale individual inputs and the output by independent positive factors. TPSR, SNIP, GenSR, and LEE modify planning, representations, or learned search (Shojaee et al., 2023; Meidani et al., 2024; Li et al., 2026; Chu et al., 2026). Appendix Z connects these choices to recent hybrid systems.

Coordinates and evaluation. Voigt et al. (2025) document numerical-domain and perturbation fragility, including limitations of standardization. Our question is which training and output choices improve final returns in a matched system. ESR relates expression equivalence, coefficient precision, and code length (Bartlett et al., 2022); MDLformer learns complexity to guide search (Yu et al., 2025). Our native-coordinate selector is a heuristic that compares stored description-length

scores in the coordinates used for refinement. Units-aware Φ-SO uses supplied physical dimensions (Tenachi et al., 2023), information absent from independent column rescaling. PySR, Operon, TIR, and ParFam provide evolutionary or continuous-family alternatives (Cranmer, 2023; Burlacu et al., 2020; de Franca, 2022; Scholl et al., 2025). SRBench and SRSD motivate broad predictive and physical evaluation (La Cava et al., 2021; Matsubara et al., 2024); adjacent neural-surrogate routes appear in Appendix H. We distinguish common-pool selection from complete systems, and numerical precision from symbolic equality (Kronberger et al., 2026).

Training and inference interfaces. Scheduled sampling studies the shift from teacher-provided to generated prefixes (Bengio et al., 2015). Unified text-to-text modeling and instruction mixtures study shared output representations and task exposure (Raffel et al., 2020; Chung et al., 2024). Our comparison crosses numerical target formats with two generation/refinement routes in one pretrained SR system. Its evidence is the resulting ranking crossover and the measured response to balanced format exposure.